UNIVERSITY OF CALGARY

Communication and Information Theory, Cryptography, and Applications

by

David Anthony Richardson

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE
DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF MATHEMATICS & STATISTICS
CALGARY, ALBERTA
September, 2006

UNIVERSITY OF CALGARY
FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled "Communication and Information Theory, Cryptography, and Applications" submitted by David Anthony Richardson in partial fulfillment of the requirements of the degree of Master of Science.

Supervisor, Dr. A. A. Bruen
Department of Mathematics & Statistics

Co-Supervisor, Dr. B. Brenken
Department of Mathematics & Statistics

Dr. B. Davies
Department of Electrical and Computer Engineering

Dr. L.M. Bates
Department of Mathematics & Statistics

Sept. 14, 2006

Date

Abstract

We present here a wide-ranging and general discussion encompassing communication channels, signal processing, and applications to cryptography using results from abstract Information Theory, a field of mathematical study invented by Claude Shannon, the famous researcher from Bell Labs. We give a brief survey of methods used to calculate channel capacity for discrete channels and use them, along with some new results, including a generalization of Shannon's regular channel theorem, to clarify some misunderstandings in the literature. A relatively novel result regarding reciprocal channels is also discussed. To our knowledge, these types of channels have not yet been fully exploited. Following a very brief discussion of concepts from signal processing, applications to cryptography are presented, beginning with the definition and proper characterization of Shannon's notion of "perfect secrecy". We conclude with some new examples of perfect secrecy and a discussion of a very interesting analog cryptosystem due to Alan Turing that is seldom discussed in the standard literature.

Acknowledgements

This thesis simply wouldn't have been possible without the assistance and dedication of my supervisor, Dr. Aiden A. Bruen. His friendship and support, both financially and academically, was much appreciated throughout my time here at the University of Calgary.

Dedication

To my wife, Meredith. You are the love of my life and without you, I'd be lost. This thesis is also dedicated to our baby on the way, who provided me with some extra motivation to complete my work!

Finally, to my family. Thank you all for your love and support!

Table of Contents

List of Figures

1 Introduction

In the Department of Mathematics and Statistics at the University of Calgary, it seems traditional for a Master's thesis to provide a survey on important results found in the literature. Our goal here is somewhat more ambitious. Not alone do we present a survey, but we also provide some new research results which have not yet appeared in print elsewhere. These new results are detailed chapter by chapter.

Our focus here will be to explore some of the insights into some of the widespread applications of abstract Information Theory, which provides a mathematical characterization of both information and communication channels, with applications ranging from data compression and cryptography to biology, the stock market, and gambling (and even jurisprudence!). We stress here that the applicability of Information Theory need not be limited to the study of electronic communication systems alone. We do mention connections with communications systems, although they are not central to our discussion.

Following the brief presentation of pertinent background material in Chapters 2 and 3, we proceed to a discussion regarding discrete channels in Chapter 4. Throughout the chapter, we correct several misunderstandings and omissions in the literature by means of examples, in the area of channel capacity. Again, such channels can be quite general. We also provide a generalization of Claude Shannon's theorem for the capacity of regular channels in the form of a new theorem, which will allow us to further clarify various examples from several sources. This new result is potentially very useful. For example, calculating the capacity of a near-regular channel with hundreds of variables reduces to a problem concerning just one variable!

Another fundamental new contribution provided by our work here involves a concept which has not been developed, namely the concept of the **reciprocal channel**. We are able to determine the capacity of the reciprocal of a binary symmetric channel by exploiting a basic

technical result, namely, the symmetry of information content, in addition to outlining several other properties of these useful channels in Chapter 5. It is important to emphasize that the calculation of the capacity of a discrete memoryless channel can sometimes be greatly simplified, as we point out here, by calculating the capacity of its reciprocal, making our development here potentially very useful in a wide range of problems. Furthermore, since the result on reciprocal channels involves the crucial symmetry of information, we provide an intuitive discussion of that result which is not given in the standard text books.

In Chapter 6, we present a very brief discussion of the sampling theorem together with some of the difficulties in applications to its practical use in the famous capacity formula for band-limited channels, $Capacity = W \log(1 + S / N)$. One difficulty in applying the capacity formula is a result of the well-known fact from the study of Fourier Transforms that, if a signal is time-limited it cannot be band-limited, and vice versa. This material is standard in the literature, so our discussion is very brief. Our main goal is to pave the way for a discussion of "analog cryptography" below.

Finally, several aspects of cryptography are explored in Chapter 7, particularly Shannon's notion of "perfect secrecy" and its proper characterization in the form of Latin squares, instead of just the so-called one-time pad. This characterization is typically not provided in the standard literature. We also provide here some new examples of perfect secrecy. To conclude our discussion, we present a novel idea using the sampling theorem for the encryption of analog signals due to Alan Turing, which shows that analog signals should not be ignored as they frequently are when discussing cryptography in general.

2 Preliminaries

Before we explore the various applications of Information Theory, it is necessary to outline some preliminary concepts, including some important results, definitions, and properties pertaining to probability theory and the study of Fourier transforms and Fourier series expansions.

2.1 Probability Theory

The study of Information Theory relies heavily upon many tools and techniques developed in the field of probability theory. By its very nature, information is inherently probabilistic: the more improbable an event is, the more information we gain upon learning that such an event has occurred. We will soon see that being able to represent information in such a way leads to some very powerful and widely applicable results. We begin with an overview of some concepts from probability theory, including the definition of a random variable.

2.1.1 Random Variables

Definition 2.1: *A* ***random variable*** *X associates a value to each of the possible outcomes contained in the sample space Ω of an experiment.*

Here, we will be concerned with discrete random variables, in which each outcome of an experiment has associated with it a probability of occurrence. In particular, we denote the probability of the i^{th} outcome of an experiment as $P(X = x_i)$ (or more briefly $P(x_i)$), which is usually assumed to be non-zero. The **probability distribution** of a random variable X is the set of probabilities

$$\{P(X = x_1), P(X = x_2), \ldots, P(X = x_n)\} = \{P(X = x)\}$$

where $\sum_{i=1}^{n} P(X = x_i) = 1$ and $0 < P(X = x_i) \leq 1$.

It must be noted that we have brushed aside the more rigorous definition of a random variable, as it is not required for our uses here. The interested reader is referred to [8] or [17] for a more formal definition.

For an example of a random variable, consider the tossing of a single coin. A suitable random variable for this experiment assumes the value of '0' for an outcome of tails and a '1' for an outcome of heads. Here, the sample space $\Omega = \{'Tails', 'Heads'\}$. Typically speaking, given the prevalence of digital information in our society today, the random variables encountered throughout our discussion will be discrete, i.e. the sample space of the associated random experiment will be countable, and most often finite in size. However, for continuous random variables, we can define a probability density function, which is analogous to a probability distribution for discrete random variables. Specifically, the **probability density** of a continuous random variable X is the function $p(x)$ where

$$P(X \leq k) = \int_{-\infty}^{k} p(x)dx \text{ and } \int_{-\infty}^{\infty} p(x)dx = 1$$

In order to compare different random variables, it is useful to determine the average or expected value of a variable, as well as the variance, which is indicative of how spread out a random variable is.

Definition 2.2: *The **expected value** of a random variable X is the weighted sum over the given probabilities of the possible values of X and is given by*

$$E(X) = \sum_{i=1}^{n} x_i P(X = x_i) \tag{2.1}$$

Definition 2.3: *The **variance** of a random variable X is the weighted average of the variable* $(x_i - E(X))^2$ *and is given by*

$$V(X) = \sum_{i=1}^{n} (x_i - E(X))^2 P(X = x_i) = E(X^2) - [E(X)]^2 \tag{2.2}$$

The variable $(x_i - E(X))^2$ provides a measure of how far each value x_i is from the average value, in terms of Euclidean distance. As such, variance is a good indicator of how spread out the values of X are. Another commonly used measure for this purpose is called the *standard deviation* of X, and is denoted $\sigma(X) = \sqrt{V(X)}$.

We can also consider pairs of random variables as a single joint random variable. If X, Y are random variables where X has possible outcomes $x_1, x_2, ..., x_n$ and Y has possible outcomes $y_1, y_2, ..., y_m$, then (X,Y) is a random variable with possible outcomes (x_i, y_j) for $1 \le i \le n$, $1 \le j \le m$, where the joint probability $P(X = x_i, Y = y_j)$ represents the probability that both $X = x_i$ and $Y = y_j$.

The individual distributions for X and Y can be obtained from the joint distribution by summing over all outcomes for X and Y respectively, i.e.

$$P(X = x_i) = \sum_{j=1}^{m} P(X = x_i, Y = y_j) \text{ and } P(Y = y_j) = \sum_{i=1}^{n} P(X = x_i, Y = y_j)$$

There are also situations where we wish to know the value of a random variable X, given that an outcome corresponding to a random variable Y has already occurred. This probability, denoted by $P(Y = y_j \mid X = x_i)$, is called the **conditional probability** of one random variable given another and plays a large role in describing communication channels, as well as in discussing security notions in cryptography. More formally, we have the following.

Definition 2.4: *Given two random variables X, Y where X has possible outcomes $x_1, x_2, ..., x_n$ and Y has possible outcomes $y_1, y_2, ..., y_m$, the* ***conditional probability*** *of X achieving the value x_i given that Y has achieved the value y_j is denoted $P(X = x_i \mid Y = y_j)$ and is defined as follows.*

$$P(X = x_i \mid Y = y_j) = \frac{P(X = x_i \text{ and } Y = y_j)}{P(Y = y_j)}$$

where it is assumed that $P(Y = y_j) \neq 0$.

Using Definition 2.4, we can now state the following formula known as **Bayes' Formula.**

Theorem 2.5 (Bayes' Formula): *Given two random variables X, Y where X has possible outcomes* $x_1, x_2, ..., x_n$ *and Y has possible outcomes* $y_1, y_2, ..., y_m$, *if* $P(X = x_i) \neq 0$ *for all* $1 \leq j \leq m$, *then*

$$P(Y = y_j \mid X = x_i) = \frac{P(X = x_i \mid Y = y_j)P(Y = y_j)}{P(X = x_i)} \qquad (2.3)$$

Note that Bayes' Formula leads us to two separate definitions for independence between two random variables and that Definition 2.6 is slightly more general in that it makes sense even in the case where both $P(X = x)$ and $P(Y = y)$ are zero.

Definition 2.6: *Two random variables* X, Y *with joint probability distribution* $\{P(X = x, Y = y)\}$ *and marginal probability distributions* $\{P(X = x)\}, \{P(Y = y)\}$ *are* ***independent*** *if*

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j) \qquad (2.4)$$

for all $1 \leq i \leq n, 1 \leq j \leq m.$

Definition 2.7: *Two random variables X, Y with joint probability distribution* $\{P(X = x, Y = y)\}$ *and marginal probability distributions* $\{P(X = x)\}, \{P(Y = y)\}$ *are* ***independent*** *if*

$P(X = x_i \mid Y = y_j) = P(X = x_i)$, or equivalently, if $P(Y = y_j \mid X = x_i) = P(Y = y_j)$.

for all $1 \leq i \leq n, 1 \leq j \leq m.$

We can also make use of Bayes' Formula along with the fact that for jointly distributed variables *X, Y,* $P(X = x_i) = \sum_{j=1}^{m} P(X = x_i, Y = y_j)$ to obtain the Law of Total Probability.

Theorem 2.8 (Law of Total Probability): *Given two jointly distributed discrete random variables X, Y where X has marginal probability distribution* $\{P(X = x)\}$*, then the marginal probabilities of Y are given by*

$$P(Y = y_j) = \sum_{i=1}^{n} P(Y = y_j \mid X = x_i)P(X = x_i) \tag{2.5}$$

for all $1 \leq i \leq n, 1 \leq j \leq m.$

Note that the same result holds for calculating the marginal probabilities of X given the joint distribution for (X,Y) and the marginal distribution for Y.

2.1.1.1 Notable Random Variables

For the purposes of studying Information Theory, there are two important types of probability distributions to consider, namely the Bernoulli distribution and the ubiquitous Gaussian, or Normal, distribution.

A Bernoulli random variable X represents the outcomes of an independently repeated experiment for which there is either a failure (with probability $1-p$) or a success (with probability p) in each trial. A random variable with this type of distribution is a **Bernoulli random variable with parameter** p. The expectation of a Bernoulli random variable with n trials is np, and the standard deviation is $np(1-p)$.

Note that Bernoulli random variables are discrete. There is also a very important continuous random variable in the background. A **Normal** or **Gaussian random variable with parameters** (μ, σ^2) is a continuous random variable with a probability density function (analogous to the distribution of a discrete random variable) given by

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \tag{2.6}$$

where μ is the **mean** and σ^2 is the **variance**. The fact that noise can be modeled using a Normal random variable with a mean of zero is important in the development of channel capacity for continuous signals.

2.1.1.2 The Weak Law of Large Numbers

Consider a sequence of n independent Bernoulli trials, where the result of a "failure" is assigned a value of '0' and the result of a "success" is assigned a value of '1'. If we observe this binary sequence long enough, what happens to the total number of 1's and 0's present in the sequence? To investigate the long-run behavior of such an experiment, we will need to make use of the Weak Law of Large Numbers: see for example [4].

Theorem 2.9 (Weak Law of Large Numbers): *Let* $X_1, X_2, X_3, \ldots$ *be a sequence of independent, identically distributed random variables with* $\mu = E(X_i)$, $\sigma^2 = V(X_i) < \infty$, $i=1,2,\ldots$. *Then* $\forall \varepsilon > 0$,

$$\lim_{n \to \infty} P\left(\left| \frac{X_1 + X_2 + \ldots + X_n}{n} - \mu \right| > \varepsilon \right) = 0 \qquad (2.7)$$

In the case of n Bernoulli trials, we point out that, here, $\mu = np$ and $p = \Pr(\text{"}Success\text{"})$.

2.2 The Fourier Transform & Fourier Series

The Fourier transform, or Fourier integral, is a method of function representation that applies to a wide range of functions, both periodic and non-periodic, and is used to represent signals in terms of frequency domain characteristics. Essentially, the Fourier transform of a time-varying function produces a frequency-varying function $H(f)$, where $|H(f)|$ is the amplitude of the function, and f represents the frequency. In general, given a time-varying function $h(t)$, assume $h(t)$ satisfies a set of conditions known as *Dirichlet conditions*, namely

i. $h(t)$ is absolutely integrable over the entire real line

ii. $h(t)$ has a finite number of maxima and minima over any finite interval of its domain

iii. $h(t)$ has a finite number of discontinuities over any finite interval of its domain.

Then $h(t)$ has a Fourier transform representation $H(f)$ where

$$H(f) = \Im\{h(t)\} = \int_{-\infty}^{\infty} h(t) e^{-2\pi i f t} dt \tag{2.8}$$

On the other hand, $h(t)$ can be recovered from $H(f)$ using the inverse Fourier transform:

$$h(t) = \Im^{-1}\{H(f)\} = \int_{-\infty}^{\infty} H(f) e^{2\pi i f t} df \tag{2.9}$$

We will see in Chapter 6 that it is also useful to represent functions in terms of their Fourier series expansion which, for a periodic function $p(t)$, is given by

$$p(t) = \sum_{n=-\infty}^{\infty} c_n e^{2\pi i n f_p t} \tag{2.10}$$

where $c_n = \frac{1}{T} \int_{\alpha}^{\alpha+T} p(t) e^{-2\pi i n f_p t} dt$ for some arbitrary α.

In addition to periodicity, the function $p(t)$ must also satisfy the Dirichlet conditions outlined above. It should be noted that any functions that can be generated by physically realizable means satisfy the Dirichlet conditions, and hence both the Fourier transform and series representations that will be utilized here are well-defined.

3 Classical Information Measures

We now focus our attention on the fundamentals of Information Theory in preparation for a look at some of the most important applications of the field. Developed in the late nineteen forties, Information Theory deals with the analysis of information from a mathematical perspective. Based upon the work of Hartley and others, Claude Shannon began his development of the theory using a logarithmic measure for information, based on the observations that:

1. It is useful in practice. Several engineering applications involve parameters that vary in proportion to the logarithm of the number of possible outcomes of a given experiment or procedure.
2. The logarithmic measure permits linearity in coding and channel capacity for example, which is what one should intuitively expect.
3. The logarithmic measure provides a great degree of simplicity in many calculations relating to Information Theory, which would not be the case if we performed the same evaluations under the framework of the number of possibilities involved.

Using the logarithmic measure, Shannon defined information to depend on the probability of the given message using the concept of **entropy**.

In addition to basic **entropy**, we will also investigate the fundamental concept of **conditional entropy**, which is a measure of redundancy between two or more random variables, or how much information they have in common. Furthermore, we will discuss **mutual information**, which is a measure of how much information a random variable reveals about another, and is fundamental in the analysis of communication channels. Finally, we briefly discuss typical sequences, which allow us to make some important conclusions about long binary sequences that have been produced by a memoryless or Markov source.

3.1 Entropy

The information gained, or equivalently, the amount of uncertainty removed, from learning the value of a random variable is referred to as the **entropy** of the random variable. Entropy is a measure of information that is based entirely upon the probability of the outcomes of a random variable, as opposed to the actual values achieved by the random variable itself.

Definition 3.1: *If X is a finite random variable with possible values $x_1, x_2, ..., x_n$ and corresponding probabilities $P(X = x_1), P(X = x_2), ..., P(X = x_n)$, then the* ***entropy*** *of X is defined as*

$$H(X) = -\sum_{i=1}^{n} P(X = x_i) \log P(X = x_i) \qquad (3.1)$$

Without loss of generality, logs will typically be considered to be base 2 and $H(X)$ is measured in Shannon bits. Again, note that in the definition, only the probabilities of the values of X are utilized, and not the actual values achieved by X.

A special case of entropy arises when $n = 2$, i.e. when X is a Bernoulli random variable with parameter p. Here, $H(X) = -p \log p - (1-p) \log(1-p) = H(p, 1-p) = H(p)$ where $H(p)$ is called the Shannon function, and is encountered frequently in the study of Information Theory as we are often dealing with sequences of Bernoulli random variables. The Shannon function, as shown in Figure 3-1, represents the amount of uncertainty removed, or the amount of information gained, from learning the outcome of a single Bernoulli experiment.

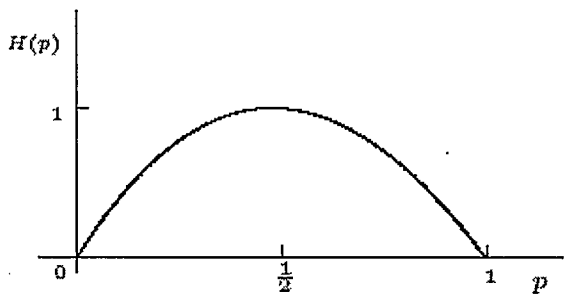

Figure 3-1 The Shannon function $H(p)$ ([2])

As shown in [4], the entropy of a random variable satisfies the following properties:

1. $H(p)$ is a concave function of p.
2. $H(X) \geq 0$, with equality if and only if $P(x_i) = 0$ or 1 for each value x_i achieved by X.
3. $H(X) \leq \log n$, where n is the number of values attained by X, with equality exactly when X has a uniform probability distribution.

Entropy plays an extremely important role in the field of **data compression.** In fact, the smallest amount of information required to describe a random variable and thus, the maximum amount a source of data that can be compressed without losing any essential information, is based entirely upon the entropy of the random variable in question.

3.2 Mutual Information, More on Entropy

As we have seen, the information gained upon learning the value of a random variable is completely dependent upon the probability distribution of the random variable. Likewise, we can define the entropy of a *pair* of jointly distributed random variables X, Y, since if X, Y are random variables with probability distributions $P\{X = x\}, P\{Y = y\}$ respectively then (X,Y) is a random variable with probability distribution $P(X = x_i, Y = y_j) = P(x_i y_j)$ for $1 \leq i \leq n, 1 \leq j \leq m$.

Definition 3.2: *If X and Y are discrete random variables where (X,Y) has possible values (x_i, y_j) with corresponding probabilities $P(x_i y_j)$, for $1 \leq i \leq n, 1 \leq j \leq m$ then the* ***joint entropy*** *$H(X,Y)$ of a pair of jointly distributed discrete random variables X, Y is defined as*

$$H(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i y_j) \log P(x_i y_j) \qquad (3.2)$$

Similarly, we can define entropy for one random variable conditioned on the value of another. Consider the random variable $(X \mid Y = y_j)$, where y_j is a fixed value achieved by the random variable Y (note that this is indeed a random variable since $\sum_{i=1}^{n} P(X = x_i \mid Y = y_j) = 1$).

Then from the definition of $H(X)$, $H(X \mid y_j) = \sum_{i=1}^{n} P(X = x_i \mid Y = y_j) \log P(X = x_i \mid Y = y_j)$.

To obtain $H(X \mid Y)$, we simply average over all values achieved by the random variable Y.

Definition 3.3: *Given two random variables X, Y, with possible values* $x_1, x_2, ..., x_n$ *and* $y_1, y_2, ..., y_m$ *respectively, the* ***conditional entropy*** *of X given Y is defined as*

$$H(X \mid Y) = \sum_{j=1}^{m} P(Y = y_j) \sum_{i=1}^{n} P(X = x_i \mid Y = y_j) \log P(X = x_i \mid Y = y_j) \qquad (3.3)$$

Alternatively, using conditional probability, we have

$$H(X \mid Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} P(X = x_i, Y = y_j) \log P(X = x_i \mid Y = y_j) \qquad (3.4)$$

where it is assumed that $P(Y = y_j) \neq 0,\ j = 1, ..., m.$

Conditional entropy is an extremely important information measure, with applications ranging from channels and their capacity, to cryptography and the concept of perfect security.

With the above definitions in place, several properties of joint and conditional entropies can now be presented. For formal proofs of the following properties, refer to [2] or [4].

Properties of Joint & Conditional Entropy:

1. $H(X,Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y) = H(Y,X)$
2. $H(X \mid Y) \neq H(Y \mid X)$ in general.
3. $H(X \mid Y) \leq H(X)$ (conditioning reduces entropy) with equality if and only if the random variables X, Y are independent, i.e. $P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$.
4. $H(f(X) \mid X) = 0$

Mutual information is another fundamental information measure. Essentially, the mutual information of two random variables is a measure of how much uncertainty is removed, or equivalently, how much information is gained from one random variable upon learning the outcome of another.

Definition 3.4: *Given two random variables X, Y, the* ***mutual information*** *of X and Y is the reduction in uncertainty of X due to the knowledge of Y and is defined as*

$$I(X;Y) = H(X) - H(X \mid Y) \tag{3.5}$$

A crucial property of mutual information that is often stated is the fact that mutual information is symmetric with respect to X and Y, i.e. $I(X;Y) = I(Y;X)$. Often lacking in the literature is an intuitive description of this property. As Welsh phrased it in [21], mutual information possesses "(a) somewhat surprising symmetry...which, as far as I can see, has no intuitive explanation".

The intuitive explanation that Welsh was looking for can be provided using *input and output fans*, and will be discussed in Chapter 4. In addition, the symmetry of mutual information for two variables can be deduced using Venn diagrams, and is shown below in Figure 3-2, following the analytic proof.

Proposition 3.5: *Given* $I(X;Y)$ *defined by* $I(X;Y) = H(X) - H(X \mid Y)$, *then* $I(X;Y) = I(Y;X)$ *for random variables X and Y.*

Proof: Let $P(X = x_i) = P(x_i)$ where X is a random variable with possible values $x_1, x_2, \ldots, x_n$ and corresponding probabilities $P(x_1), P(x_2), \ldots, P(x_n)$ and similarly, let $P(Y = y_j) = P(y_j)$ where Y is a random variable with possible values $y_1, y_2, \ldots, y_m$ and corresponding probabilities $P(y_1), P(y_2), \ldots, P(y_m)$.

Since $I(X;Y) = H(X) - H(X \mid Y)$ by definition, we have

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X \mid Y) \\
&= -\left(\sum_{i=1}^{n} P(x_i)\log(P(x_i))\right) - \left(-\sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i y_j)\log(P(x_i \mid y_j))\right) \\
&= -\sum_{i=1}^{n} P(x_i)\log(P(x_i)) + \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i y_j)\log(P(x_i \mid y_j)) \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i y_j)\log(P(x_i)) + \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i y_j)\log(P(x_i \mid y_j)) \\
&= -\left(\sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i y_j)\log(P(x_i)) - \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i y_j)\log(P(x_i \mid y_j))\right) \\
&= -\left(\sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i y_j)\log\left(\frac{P(x_i)}{P(x_i \mid y_j)}\right)\right)
\end{aligned}
$$

Now, from Bayes' formula, we have $P(x_i \mid y_j) = \dfrac{P(x_i)\cdot P(y_j \mid x_i)}{P(y_j)}$, so that

$$
\begin{aligned}
I(X;Y) &= -\left(\sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i y_j)\log\left(\frac{P(x_i)\cdot P(y_j)}{P(x_i)\cdot P(y_j \mid x_i)}\right)\right) \\
&= -\left(\sum_{i=1}^{n}\sum_{j=1}^{m} P(y_j x_i)\log(P(y_j)) - \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i y_j)\log(P(y_j \mid x_i))\right)
\end{aligned}
$$

$$= -\sum_{j=1}^{m} P(y_j)\log\left(P(y_j)\right) - \left(-\sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i y_j)\log\left(P(y_j \mid x_i)\right)\right)$$

$$= H(Y) - H(Y \mid X) = I(Y;X)$$

□

As a result of the symmetry of mutual information, we have that

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) = I(Y;X) \tag{3.6}$$

The various relationships between entropy, joint entropy, conditional entropy, and mutual information can be summed up in the Venn diagram taken from [2] shown in Figure 3-2.

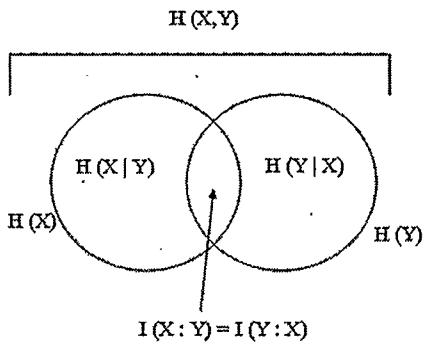

Figure 3-2 The relationship between entropy and mutual information

Remark: The analogous Venn diagram for three variables does not provide the same result.

3.3 Typical Sequences

Consider a source emitting n consecutive binary digits independently and according to identical distributions (i.e. the source is *i.i.d.*). This process represents a sequence of n independent Bernoulli trials, with parameter p. According to the Weak Law of Large Numbers, the number of 1's and 0's present in the sequence of bits emitted from the source approach, in terms of probability, the expected value of each, namely np and $n(1-p)$ as n gets larger. Hence, "most" of the time, a sequence of length n will have close to the average number of 1's and 0's. More specifically, from [2], if we choose an $\varepsilon > 0$ and any integer k such that $k^2 < 2/\varepsilon$, then with N = actual number of 1's in the sequence, we say that the sequence is **typical** if $\left|\frac{N-np}{\sqrt{np(1-p)}}\right| < k$. It will then follow that the set of non-typical sequences of length n has a total probability less than ε.

More generally, for an *i.i.d.* source, we consider typical sequences as sequences that are "supposed" to occur. Since it is highly unlikely that we would get a run of n 0's or n 1's for example, these sequences, called *atypical sequences*, are disregarded when analyzing the statistical behaviour of a communications system. Essentially, the set of typical sequences account for the majority of the probability and the minority of the number of sequences that are emitted from an *i.i.d.* source. The following theorem adapted from [4] summarizes the most important properties of typical sequences, where $H(X)$ is the entropy of the random variable X.

Theorem 3.6 (Typical Sequences): *If X is a random variable with possible values $\{x_1, x_2, ..., x_m\}$ and entropy $H(X)$, then with $A_\varepsilon^{(n)}$ representing the set of typical sequences of length n, the following hold:*

1. *If* $(x_1, x_2, ..., x_m) \in A_\varepsilon^{(n)}$ *then* $\left|-\frac{1}{n}\log P(x_1, x_2, ..., x_m)\right| \in \left(H(X)-\varepsilon, H(X)+\varepsilon\right)$

2. $P((x_1, x_2, ..., x_m) \in A_\varepsilon^{(n)}) > 1 - \varepsilon$ *for n sufficiently large.*
3. $\left|A_\varepsilon^{(n)}\right| \leq 2^{n(H(X)+\varepsilon)}$
4. $\left|A_\varepsilon^{(n)}\right| \geq (1-\varepsilon)2^{n(H(X)-\varepsilon)}$

To paraphrase Theorem 3.6, the probability of any sequence from the typical set is almost uniform, the probability that a given sequence is contained in the typical set is close to 1, and the total number of typical sequences is approximately $2^{nH(X)}$. The idea of typical sequences will be called upon to aid in the development of Shannon's capacity theorem for discrete channels.

With the preliminaries out of the way, we can now proceed with the investigation and analysis of some of the most important applications of Information Theory, particularly channels and their capacity, signal processing, and cryptography.

4 Discrete Channels & Their Capacity

One of the most fundamental results due to Claude Shannon pertains to the maximum rate of information that can be transmitted over an inherently noisy electronic channel, such that the inevitable errors due to interference or noise encountered by the message can be made arbitrarily small. This limit to effective communications is referred to as the **capacity** of the channel, where a channel can be thought of as a mechanism for transmitting messages. While this result may seem somewhat subtle at first, it should be noted that the entire communications engineering community at the time believed that such a reduction in signal errors was possible only if the information rate approached zero. Hence, Shannon's results, dating from about 1940, almost single-handedly revolutionized the field of efficient communication system design. In fact, to this day, most of the practical work in communications still revolves around Shannon's work.

In this chapter, we will focus on the case where the channels are discrete. The continuous version of Shannon's capacity theorem will be presented in Chapter 6. A discussion of Shannon's capacity theorem for discrete memoryless channels will be presented, followed by an overview and analysis of a variety of channel types. We also present, in effect, a brief survey of available methods for calculating the capacity of a wide range of channels, in addition to providing several new results which will further aid us in the task of evaluating discrete channels. We will then use these methods and results to address discrepancies and obscurities that appear in various published sources pertaining to the calculation of capacity.

4.1 Discrete Channels

Definition 4.1: *A discrete memoryless channel (DMC) is a mapping* $\Gamma : \mathrm{X} \rightarrow \mathrm{Y}$ *where* X *is the set of values attained by the random variable X and* Y *is the set of values attained by the random variable Y, both of which are discrete and of finite size.*

Every DMC has associated with it a **channel transition matrix** $P = [P(Y = y \mid X = x)]$. From this point forward, as a convention, we will be using the notation '$x \in X$' to denote that x is a value achieved by the random variable X. Again, for our purposes here, we will assume that any given channel is **memoryless**. Essentially, this condition implies that each input that is passed through the channel does not depend on any of the inputs that preceded it.

From an abstract point of view, a channel is a mechanism that is used for converting inputs into outputs in a probabilistic sense. In fact, it can be shown that all that is required to define a discrete memoryless channel is a joint probability distribution and vice versa.

Theorem 4.2 (Discrete Channels <=> Joint Probability Distribution): *Given a discrete channel* $\Gamma : X \to Y$, *there exists a joint probability distribution* $P(X = x, Y = y)$ *and, conversely, given a joint probability distribution* $P(X = x, Y = y)$, *one can construct a discrete channel* $\Gamma : X \to Y$.

Proof: By definition, if we have a discrete channel Γ, then the channel transition probabilities given by $P(Y = y \mid X = x)$ are known. Then, using Bayes' formula, with $P(Y = y) > 0$ for all $y \in Y$, we have

$$P(Y = y \mid X = x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{P(X = x, Y = y)}{\sum_{y \in Y} P(X = x, Y = y)} \qquad (4.1)$$

Therefore, we can obtain $P(X = x, Y = y)$ for all $x \in X$, $y \in Y$, so that having a channel implies having a joint probability distribution.

Alternatively, given a joint probability distribution, we can determine the channel transition probabilities and so a channel matrix, and thus obtain a channel, since

$$\frac{P(X = x, Y = y)}{\sum_{y \in Y} P(X = x, Y = y)} = \frac{P(X = x, Y = y)}{P(X = x)} = P(Y = y \mid X = x) \qquad (4.2)$$

with $P(X = x) > 0$ for all $x \in X$ and $\sum_{y \in Y} P(X = x, Y = y) = P(X = x)$. Hence, having a channel implies possessing a joint probability distribution between the inputs and outputs, and vice versa.

□

With the definition of mutual information and discrete channels firmly in hand, we can now define the **capacity** of a discrete channel, which can be thought of as the greatest rate of information that can be transmitted such that the probability of erroneous transmission can be made arbitrarily small.

Definition 4.3 (Capacity of a Discrete Channel): *The* ***capacity*** *of a discrete channel* $\Gamma : X \to Y$ *is defined as* $\Lambda = \max_{P(X=x)} I(X;Y)$*, where the maximum is taken over all possible input distributions given by* $\{P(X = x)\}$*. Equivalently, the capacity of a discrete channel can also be defined as* $\Lambda = \max_{P(Y=y)} I(Y;X)$

An informal proof of Shannon's capacity theorem for certain discrete channels will be provided shortly. Some of the most important properties of channel capacity, taken from [17], are summarized below.

1. $\Lambda \geq 0$, since $I(X;Y) \geq 0$
2. $\Lambda = \log|$ # of distinguishable inputs$|$
3. $\Lambda = \log|$ # of distinguishable outputs$|$
4. $I(X;Y)$ is a continuous function of $P(X = x)$
5. $I(X;Y)$ is a concave function with respect to $P(X = x)$

More formally, the capacity is defined as the supremum of $I(X;Y)$ over all input distributions, but $I(X;Y)$ is both a continuous and concave function of $P(X = x)$, so that a local maximum is in fact a global maximum. Then, since I is concave over a closed convex set, the function does indeed attain the maximum.

4.1.1 Capacity of a Binary Channel

We will now present an interpretive sketch of the capacity theorem for (n parallel copies of) a binary symmetric channel, which will allow us to proceed with a more intuitive grasp of channel capacity. For a detailed, more formal proof of Shannon's famous result, the reader is referred to [19].

Consider a source X emitting sequences of binary digits of length n. We are interested in determining the greatest number of distinguishable inputs emitted by X that can be transmitted over an arbitrary discrete channel subject to random noise.

From Chapter 3, we know that since X is a random variable with a corresponding probability distribution $\{P(X = x)\}$, then there are roughly $2^{nH(X)}$ typical input sequences of length n. Similarly, at the output, there are approximately $2^{nH(Y)}$ typical output sequences of length n since Y can be regarded as a source (using the fact that $P(Y = y) = \sum_{y \in Y} P(Y = y \mid X = x)P(X = x)$).

If we fix an input sequence **x** and transmit the sequence, a natural question to ask is where may **x** end up upon transmission over the noisy channel? In other words, following transmission, what output sequences could have reasonably been caused by **x**? The answer lies in the fact that all output sequences are "noisy" versions of input sequences, where the noise can act on **x** to produce one of $2^{nH(Y|X)}$ probable output sequences. For some insight into why this is indeed the case, recall that $H(Y \mid x)$ is a random variable and $H(Y \mid X)$ is obtained by averaging $H(Y \mid x)$ over all inputs x, and that each possible output sequence occurs with the same probability from our definition of typical sequences in Chapter 3. Therefore, the input sequence **x** can be mapped to one of $2^{nH(Y|X)}$ output sequences, producing an "output fan" of likely sequences as shown in Figure 4-1 below.

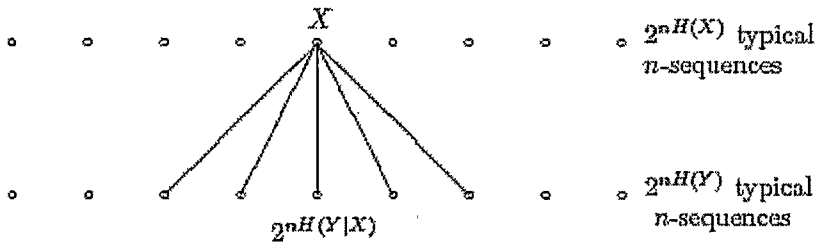

Figure 4-1 Output fan associated with the input sequence x

Upon receiving an output sequence contained in a given output fan, we must be able to unambiguously determine which input message was sent. In order to do so, we require that no two output fans can overlap since, if they did, then given an output sequence **y** in a particular output fan, we would not be able to say with certainty which input sequence produced the received output sequence. This "non-overlap" condition can be expressed mathematically as

$$N2^{nH(Y|X)} \leq 2^{nH(Y)} \tag{4.3}$$

where N represents the total number of output fans. Expressing the above in terms of N, we have

$$N \leq 2^{n(H(Y)-H(Y|X))} = 2^{nI(Y;X)} \tag{4.4}$$

Upon taking the logarithm of both sides and dividing through by n, we obtain

$$\frac{\log N}{n} \leq I(Y;X) \tag{4.5}$$

Finally, maximizing over all output distributions for Y, we obtain the fundamental result

$$\frac{\log N}{n} = \Lambda \tag{4.6}$$

where the left hand side is thought of as the **transmission rate** in bits per symbol since, in the binary case, there are log N message bits, and n is the number of symbols per message string.

On the other hand, consider the situation depicted below in Figure 4-2. In this case, we fix an output symbol and attempt to determine the set of inputs that could have produced it via transmission over a noisy channel. Again, we have about $2^{nH(X)}$ typical input sequences and approximately $2^{nH(Y)}$ typical output sequences, all of length n.

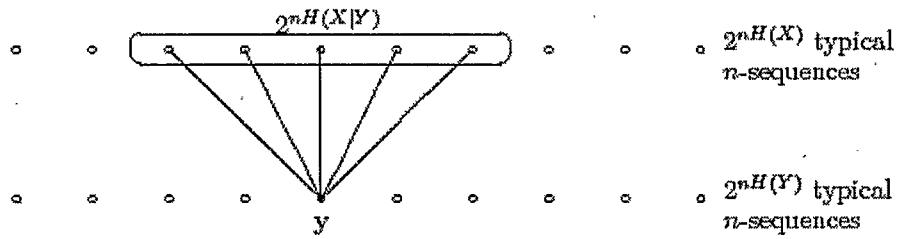

Figure 4-2 Input fan associated with the output sequence y

In this case, given a received output sequence **y**, the question becomes 'which input sequence(s) could have most likely produced the output sequence **y**?' In this case, we can view our channel as transmitting from Y to X, since we can regard Y as a source (so-called "flipped", or "reciprocal" channels will be discussed in detail in Chapter 5). Therefore, by the same argument as before, the output sequence **y** can be mapped to one of $2^{nH(X|Y)}$ input sequences, each occurring with equal probability, thus producing an input fan of likely sequences as shown in Figure 4-2. In order to ensure unambiguous communication, we must again have the "non-overlapping" condition which, in this case, is expressed as

$$M2^{nH(X|Y)} \leq 2^{nH(X)} \tag{4.7}$$

where M represents the total number of input fans. Therefore, in terms of M, we have

$$M \leq 2^{nH(X)} - 2^{nH(X|Y)} = 2^{nI(X;Y)} \tag{4.8}$$

and upon taking the base 2 logarithm of both sides and dividing through by n, we have

$$I(X;Y) \geq \frac{\log M}{n} \tag{4.9}$$

This time, by maximizing over all possible *input* distributions for X, we obtain

$$\Lambda = \frac{\log M}{n} \tag{4.10}$$

We therefore conclude that the capacity of n parallel copies of an arbitrary, discrete binary symmetric channel is given by

$$\Lambda = \max_{P(X=x)} I(X;Y) = \max_{P(Y=y)} I(Y;X)$$

It is also worth mentioning that we have seen yet another justification for the fact that mutual information is indeed symmetric with respect to a source and a destination. This symmetry will prove to be extremely useful when examining reciprocal channels in Chapter 5.

4.2 Binary Channels

A type of channel that is of particular practical importance is the binary symmetric channel, given the prevalence of binary data in most practical computing applications. We will first investigate the Binary Symmetric Channel, or BSC which, by definition, features a probability of bit error that is the same for either input bit value. The general binary channel will also be discussed in this section, where since the probability of bit error varies for each input bit value, computation of channel capacity is somewhat more involved.

4.2.1 The Binary Symmetric Channel

The Binary Symmetric Channel (BSC) consists of an input and output alphabet $\{0,1\}$ along with a parameter p, which represents the probability of a bit error during transmission, as depicted in Figure 4-3.

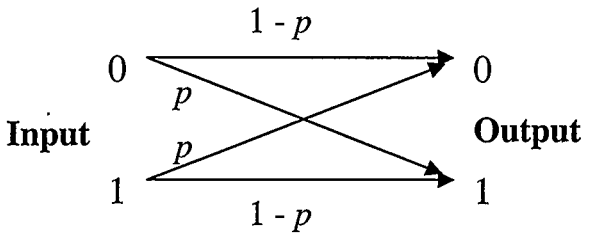

Figure 4-3 A binary symmetric channel with parameter p

The channel transition matrix of a BSC is given by

$$\mathrm{P} = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}$$

If the parameter p is equal to ½ , the channel is completely unreliable, since any input bit is capable of producing the output '0' or '1' with equal probability. Thus, the receiver has no way of knowing which input bit was sent. It should also be noted that if $p > ½$, we can switch the inputs '0' and '1', resulting in the original BSC with parameter $1-p$. *Hence, we will only consider parameter values p such that* $0 \leq p \leq ½$. We can now formally present Shannon's capacity theorem for the BSC.

Theorem 4.4 (Capacity of a BSC): *The capacity of a BSC with parameter p is given by* $\Lambda = 1 - H(p)$ *where* $H(p)$ *is the Shannon function.*

Proof: Let $P(X=0) = \alpha, P(X=1) = 1-\alpha$. Now, the output symbol will be '0' if either '0' is input and transmitted without error, or if '1' is input and is erroneously transmitted.

Thus, $P(Y=0)=\alpha(1-p)+(1-\alpha)p$, and using similar arguments, $P(Y=1)=\alpha p+(1-\alpha)(1-p)$.

From the channel matrix P, we can compute

$$H(Y\mid X)=-\sum_{x\in X}P(X=x)\sum_{y\in Y}P(Y=y\mid X=x)\log P(Y=y\mid X=x)$$

$$=-P(X=0)P(Y=0\mid X=0)\log P(Y=0\mid X=0)-\ldots-P(X=1)P(Y=1\mid X=1)\log P(Y=1\mid X=1)$$

$$=-\alpha(1-p)\log(1-p)-\alpha p\log p-(1-\alpha)p\log p-(1-\alpha)(1-p)\log(1-p)=\big(\alpha+(1-\alpha)\big)H(p)=H(p)$$

Then, using the definition for mutual information, we have

$$I(X;Y)=I(Y;X)=H(Y)-H(Y\mid X)$$
$$=-(\alpha(1-p)+(1-\alpha)p)\log(\alpha(1-p)+(1-\alpha)p)-(\alpha p+(1-\alpha)(1-p))\log(\alpha p+(1-\alpha)(1-p))-H(p)$$

Note that $I(X;Y)$ is a function of one variable and a parameter p, which is considered fixed. To determine the maximum value of $I(X;Y)$, we need only evaluate the endpoints of the interval over which $I(X;Y)$ is defined (namely the closed interval [0,1]), or at any point where the derivative of $I(X;Y)$ is 0.

At the endpoints, if $\alpha=0$, we have $I(X;Y)=-p\log p-(1-p)\log(1-p)-H(p)=0$.
On the other hand, if $\alpha=1$, we have $I(X;Y)=-(1-p)\log(1-p)-p\log p-H(p)=0$.

Thus, in order to determine the local and, subsequently, the global maximum, it suffices to calculate the value of α for which the derivative of $I(X;Y)$ is 0. Then, with $f(\alpha)=I(X;Y)$, we obtain

$$f'(\alpha)=-(1-2p)\log(\alpha(1-p)+(1-\alpha)p)-(1-2p)-(2p-1)\log(\alpha p+(1-\alpha)(1-p))-(2p-1)$$

$$=(2p-1)\big(\log(\alpha(1-p)+(1-\alpha)p)-\log(\alpha p+(1-\alpha)(1-p))\big)$$

Therefore, $f'(x) = 0 \Rightarrow \log(\alpha(1-p)+(1-\alpha)p) = \log(\alpha p+(1-\alpha)(1-p)) \Rightarrow \alpha = \frac{1}{2}$.

Hence, the maximum value of $I(X;Y)$ is attained when $\alpha = ½$, and thus the capacity of the BSC with parameter p is given by

$$\Lambda = \max I(X;Y) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} - H(p) = 1 - H(p)$$

□

Evidently as can be seen from the above, even for one of the most straightforward channels, calculating the corresponding capacity can be quite tedious. However, later on we provide a much simpler proof of Theorem 4.4 using an observation of Shannon.

4.2.2 General Binary Channels

As an alternative to the BSC, we can consider the case in which the probability of a bit error is not the same for each input bit.

Definition 4.5: *A* ***general binary channel*** *is defined as a map* $\Gamma : X \to Y$, *where* $X, Y = \{0,1\}$ *with corresponding channel transition matrix given by*

$$P = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}$$

where p_{ij} *represents the probability that given message i was sent, message j was received, with the constraint that* $p_{00} + p_{01} = p_{10} + p_{11} = 1$. To compute the capacity of general binary channels, we can make use of a theorem due to Ash ([1]).

Theorem 4.6 (Capacity of a Binary Channel): *Consider a general binary channel with channel transition matrix*

$$\mathrm{P} = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}$$

The capacity of this channel is given by $\Lambda = \log\left(2^{u} + 2^{v}\right)$ *where*

$$u = \frac{p_{11}H(p_{00}) - p_{01}H(p_{10})}{p_{10} - p_{00}} \text{ and } v = \frac{-p_{10}H(p_{00}) + p_{00}H(p_{10})}{p_{10} - p_{00}}$$

For example, if $p_{00} = p_{11} = 1-p$, $p_{10} = p_{01} = p$ then

$$u = \frac{(1-p)H(p) - pH(p)}{p-(1-p)} = -H(p) \text{ and } v = \frac{-pH(p) + (1-p)H(p)}{p-(1-p)} = -H(p)$$

which results in a capacity of

$$\Lambda = \log\left(2^{-H(p)} + 2^{-H(p)}\right) = \log\left(2 \cdot 2^{-H(p)}\right) = \log 2 + \log 2^{-H(p)} = 1 - H(p)$$

This is exactly what we expected to obtain, since this case corresponds to the BSC discussed in Section 4.2.1. As a numerical example, consider the general binary channel with channel transition matrix

$$\mathrm{P} = \begin{bmatrix} 0.75 & 0.25 \\ 0.15 & 0.85 \end{bmatrix}$$

Then, with

$$u = \frac{(0.85)H(0.75) - (0.25)H(0.15)}{0.15 - 0.75} \approx -0.90, \; v = \frac{-(0.15)H(0.75) + (0.75)H(0.15)}{0.15 - 0.75} \approx -0.56$$

the capacity is $\Lambda = \log\left(2^{-0.90} + 2^{-0.56}\right) \approx 0.28$.

4.3 Further Techniques for Calculating the Capacity of a Discrete Channel

We now present additional techniques that will prove to be quite useful in the calculation of capacities for a variety of channels.

4.3.1 Regular Channels

Consider the channel shown in Figure 4-4.

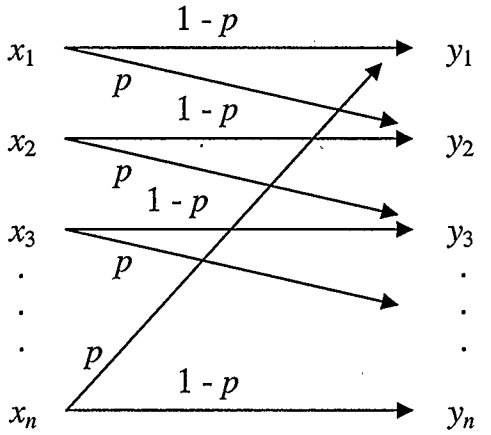

Figure 4-4 An example of a regular channel

The inherent symmetry of this type of channel can be exploited to greatly simplify the computation of its associated capacity. From a combinatorial standpoint, notice that each $x \in X$ has the same set of probabilities on the emerging lines, or in other words, their output fans look the same. This implies that, for all $x \in X$, $P(Y = y \mid X = x)$ is constant with respect to the inputs. *Let us now assume that the output fans are in fact the same.* This is tantamount to assuming that the rows of the channel matrix are permutations of each other. The channel matrix corresponding to the channel in Figure 4-4 is then called a **semi-regular** channel matrix. For example, the channel matrix P of Figure 4-4 is semi-regular with

$$\mathrm{P}=\begin{bmatrix} 1-p & p & 0 & \cdots & 0 \\ 0 & 1-p & p & \cdots & 0 \\ 0 & 0 & 1-p & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p & 0 & 0 & \cdots & 1-p \end{bmatrix}$$

Recall that

$$H(Y \mid X) = -\sum_{x \in X} P(X = x) \sum_{y \in Y} P(Y = y \mid X = x) \log P(Y = y \mid X = x)$$

Then, since the output fans are equal and the sum of the probabilities for X add up to 1, we obtain the following result.

Result 1: *The capacity* Λ *of a channel for which the rows are permutations of each other is given by the following formula.*

$$\Lambda = \max_{P(Y=y)} \{H(Y)\} - H(\textit{any row of } \mathrm{P}) \qquad (4.11)$$

In the diagrammed example above, this gives $\Lambda = \max_{P(Y=y)} \{H(Y)\} - H(p)$.

Next, let us also assume that the columns of the channel matrix are permutations of each other. Thus, combinatorially, the input fans of each $y \in Y$ are also the same, and the channel is referred to in this case as a **regular channel.**

It follows that if the input probabilities $P(X = x)$ are equal, then Y also has the equiprobable distribution so that $H(Y) = \log(\textit{number of outputs})$. From Result 1, we then have the following.

Theorem 4.7 (Capacity of a Regular Channel): *Given a regular channel* $\Gamma : X \to Y$ *with inputs* $X = \{x_1,...,x_m\}$ *and outputs* $Y = \{y_1,...,y_n\}$, *let* $(p_1, p_2,..., p_n)$ *denote any row of the channel matrix. Then the capacity* Λ *of the channel is given by the following formula.*

$$\Lambda = \log n - H(p_1, p_2,..., p_n) \tag{4.12}$$

Since the BSC is a regular channel, we now have the easier proof for the capacity of the BSC promised earlier.

Corollary 4.8: *The capacity* Λ *of the Binary Symmetric Channel with parameter p is equal to*

$$\Lambda = \log n - H(p_1, p_2,..., p_n) = 1 - H(p) \tag{4.13}$$

Remark 1: This formula easily extends to the case of *n* copies in parallel of the BSC.

Remark 2: In order to achieve capacity, we can take the *X* to be equiprobable so that the distribution for *Y* is equiprobable. However, there exist examples where the *Y* can be made equiprobable without the *X* being equiprobable. Here is one such example due to Shannon.

Example 4.9: Consider the channel in Figure 4-5 with channel transition matrix

$$P = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1/2 & 1/2 \\ 1/2 & 0 & 0 & 1/2 \end{bmatrix}$$

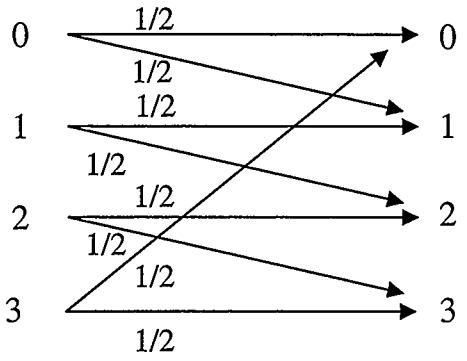

Figure 4-5 A regular channel with four inputs and outputs

Then, by Theorem 4.7, the capacity is given by $\Lambda = \log 4 - \log 2 = 1$.

Now, since the channel is regular, we know that $H(Y \mid X)$ is constant with respect to each input, so that achieving capacity amounts to maximizing $H(Y)$ with respect to all possible output distributions. From Chapter 3, we know that $H(Y)$ is maximized if and only if the outputs are equiprobable. Therefore, for each $y \in Y$, $P(Y = y) = 1/4$. If we denote the input probabilities as $P(X = 0) = \alpha, P(X = 1) = \beta, P(X = 2) = \gamma$, and $P(X = 3) = \delta$ where $\alpha + \beta + \gamma + \delta = 1$, then using the Law of Total Probability, we obtain the following system of equations relating the input probabilities to the equiprobable outputs:

$$\begin{bmatrix} 1/2 & 1/2 & 0 & 0 & 1/4 \\ 0 & 1/2 & 1/2 & 0 & 1/4 \\ 0 & 0 & 1/2 & 1/2 & 1/4 \\ 1/2 & 0 & 0 & 1/2 & 1/4 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 0 & 0 & 1 & 1/2 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & 1 & 1/2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

where '$\Rightarrow$' denotes row-reduction.

Therefore, we obtain a solution of

$$X = \begin{bmatrix} P(X=0) \\ P(X=1) \\ P(X=2) \\ P(X=3) \end{bmatrix} = \begin{bmatrix} 1/2 \\ 0 \\ 1/2 \\ 0 \end{bmatrix} + t \begin{bmatrix} -1 \\ 1 \\ -1 \\ 1 \end{bmatrix}, \text{ where } 0 \le P(X = x) \le 1 \text{ for all } x \in X.$$

Thus, we actually have **an infinite number of possibilities** such that the outputs are equiprobable but the inputs are not. In other words, there are infinitely many input probability distributions achieving capacity! For example, if $t = 0$, we can achieve capacity using only the 1st and 3rd symbols, so that the input distribution is given by $\{1/2,0,1/2,0\}$, which produces a uniform output distribution. For the other extreme, we could have used only the 2nd and 4th symbols, which corresponds to an input distribution of $\{0,1/2,0,1/2\}$, again resulting in an equiprobable output distribution.

The technique involved in analyzing regular channels can also be extended to include channels that are 'almost' regular, i.e. channels that are regular with the exception of one input and one output. This result is based upon the subtle pooling inequality, which essentially states that, for entropy, "the more equal the probabilities, the bigger the entropy." ([8]) The pooling inequality does not seem to follow from the usual Jensen's inequality argument.

We now present an extension to Theorem 4.7 that applies to the case where a channel is near-regular, i.e. the input fans are the same with the exception of one output, and the set of output fans are the same with the exception of one input.

Theorem 4.10 (Near-Regular Channel Capacity): *Let* $\Gamma : X \to Y$ *be a channel with inputs* $X = \{x_1,...,x_m\}$ *and outputs* $Y = \{y_1,...,y_n\}$. *Assume the following.*

1. *Each of* $\{x_2,...,x_m\}$ *has the same output fan*
2. *Each of* $\{y_2,...,y_n\}$ *has the same input fan*
3. $P(Y = y_1 \mid X = x_i) = P(Y = y_1 \mid X = x_j)$ *for* $2 \le i, j \le m$.
4. $P(X = x_1 \mid Y = y_i) = P(X = x_1 \mid Y = y_j)$ *for* $2 \le i, j \le n$.

Then at capacity, we must have $P(Y = y_2) = ... = P(Y = y_n)$. *This can be achieved by putting* $P(X = x_2) = ... = P(X = x_m)$.

Notice that if n and m were 100 for example, then we could assume at the outset that the last 99 inputs have the same probability distribution. Hence, it is possible to reduce an optimization problem involving 100 variables to one involving only a single variable!

For an application of Theorem 4.10, consider the channel below in Figure 4-6 discussed in Shannon [19].

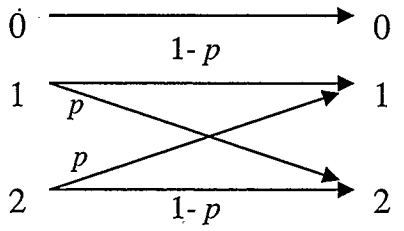

Figure 4-6 Example of an almost regular channel from [19]

Following Shannon's notation, we denote the input probabilities by P, Q and R for 0, 1, and 2 respectively. The author assumes that $Q = R$. Following this, he calculates the capacity, with the assumption that $Q = R$ simplifying the calculations. We want to point out that Theorem 4.10 provides one way of showing that we may assume $Q = R$, since the inputs have the same output fan with the exception of the input '0', and the outputs have the same input fans with the exception of the output '0'. It should also be noted that for this problem, Shannon uses the method of Lagrange Multipliers discussed below, in addition to assuming that the channel matrix P has an inverse. This assumption is too stringent, since it excludes channels where the number of inputs and the number of outputs are different. In fact, Shannon computes the capacities of such channels in [19] without making use of this assumption at all.

4.3.2 The Method of Lagrange Multipliers

The method of Lagrange Multipliers is a standard tool in solving constrained optimization problems and, as such, it will be outlined here. This method is based upon a theorem due to Lagrange, which specifies the conditions for which there exists a maximum or minimum value for a so-called *objective function*, subject to a *constraint function*. Lagrange's theorem,

slightly modified from Larson *et al* ([12]) to account for three variables and provided without proof, can be summarized as follows.

Theorem 4.11 (Lagrange's Theorem): *Let f, g have continuous first partial derivatives such that* $f(\alpha,\beta,\gamma)$ *has an extremum at a point* $(\alpha_0,\beta_0,\gamma_0)$ *on the smooth constraint curve* $g(\alpha,\beta,\gamma)=c$. *If* $\nabla g(\alpha_0,\beta_0,\gamma_0)\neq 0$, *then there is a real number* λ *such that*

$$\nabla f(\alpha_0,\beta_0,\gamma_0)=\lambda\nabla g(\alpha_0,\beta_0,\gamma_0) \tag{4.14}$$

Theorem 4.12 (The Method of Lagrange Multipliers): *Let f, g satisfy the hypothesis of Theorem 4.11, and let* $f(\alpha,\beta,\gamma)=I(X;Y)$ *have a maximum subject to the constraint function* $g(\alpha,\beta,\gamma)=\alpha+\beta+\gamma-1=0$. *Then the following steps will yield the maximum value of* $f(\alpha,\beta,\gamma)=I(X;Y)$ *and hence the capacity of the corresponding discrete channel.*

1. *Solve the system of equations* $\begin{cases}\nabla f(\alpha_0,\beta_0,\gamma_0)=\lambda\nabla g(\alpha_0,\beta_0,\gamma_0)\\ g(\alpha,\beta,\gamma)=\alpha+\beta+\gamma-1=0\end{cases}$

2. *Evaluate* $f(\alpha,\beta,\gamma)=I(X;Y)$ *at each point obtained from step 1. The largest value attained by* $I(X;Y)$ *subject to* $g(\alpha,\beta,\gamma)=\alpha+\beta+\gamma-1=0$ *is the capacity of the channel.*

Remark: For our purposes, the numbers α,β,γ always turn out to be probabilities.

With an arsenal of techniques for computing capacity in hand, we can now address some capacity questions that appear in Goldie & Pinch ([8]) and correct the results there.

Example 4.13: Consider the ternary channel shown in Figure 4-7 below.

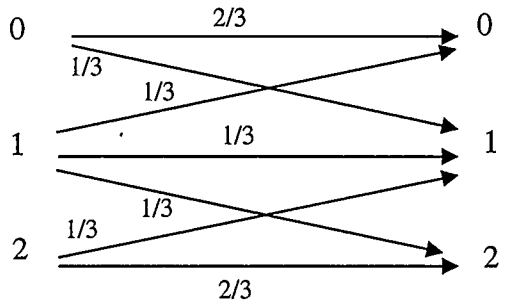

Figure 4-7 Ternary channel with one random input

The corresponding channel transition matrix is then

$$P = \begin{bmatrix} \frac{2}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

To compute the capacity of this channel, we must maximize the mutual information of the channel over all input (or output) probability distributions. However, we will seek to maximize mutual information in terms of probability p, and then set p equal to 1/3 in order to obtain the result we desire.

Therefore, the channel transition matrix that we are interested in is given by

$$P := \begin{bmatrix} 2p & p & 0 \\ p & p & p \\ 0 & p & 2p \end{bmatrix}$$

First, we can calculate $H(Y)$ by regarding Y as a source using the Law of Total Probability, where the input probability distribution is $\{P(X = x)\} = \{\alpha, \beta, \gamma = 1 - \alpha - \beta\}$:

$$P(Y=0)=\sum_{x\in X}P(Y=0\mid X=x)P(X=x)=P(Y=0\mid X=0)P(X=0)+\ldots+P(Y=0\mid X=2)P(X=1)$$

$$=2p\alpha+p\beta$$

Similarly, we compute $P(Y=1)=(\alpha+\beta+\gamma)p, P(Y=2)=p\beta+2p\gamma$. Thus we have

$$H(Y)=-(2\alpha p+\beta p)\log(2\alpha p+\beta p)-((\alpha+\beta+\gamma)p)\log((\alpha+\beta+\gamma)p)-(\beta p+2\gamma p)\log(\beta p+2\gamma p)$$

Next, we must determine $H(Y\mid X)$ and find the maximum value of $I(Y;X)=H(Y)-H(Y\mid X)$.

By definition, we have that $H(Y\mid X)=\sum_{x\in X}P(X=x)\sum_{y\in Y}P(Y=y\mid X=x)\log P(Y=y\mid X=x)$.

Then, using the entries of P, we have

$$H(Y\mid X)=-P(X=0)P(Y=0\mid X=0)\log P(Y=0\mid X=0)-\ldots-P(X=2)P(Y=2\mid X=2)\log P(Y=2\mid X=2)$$

$$=-\alpha(2p\log(2p)+p\log(p))-\beta(3p\log(p))-\gamma(2p\log(2p)+p\log(p))$$

Now, to find the capacity of the channel, we must maximize $I(Y;X)=H(Y)-H(Y\mid X)$ over all possible output probability distributions. Note that $I(Y;X)$ is a function of three variables in addition to p, subject to the constraint $\alpha+\beta+\gamma=1$. Hence, we can use the method of Lagrange Multipliers to determine the capacity, where the objective function is $f(\alpha,\beta,\gamma)=H(Y)-H(Y\mid X)$ and the constraint function is given by $g(\alpha,\beta,\gamma)=\alpha+\beta+\gamma-1$.

Computing partial derivatives, we have:

$$f_\alpha(\alpha,\beta,\gamma)=-2p\log(2\alpha p+\beta p)-p\log((\alpha+\beta+\gamma)p)-3p+2p\log 2p+p\log p$$

$$f_\beta(\alpha,\beta,\gamma)=-p\log(2\alpha p+\beta p)-p\log((\alpha+\beta+\gamma)p)-p\log(\beta p+2\gamma p)-3p-3p\log p$$

$$f_\gamma(\alpha,\beta,\gamma)=-2p\log(\beta p+2\gamma p)-p\log((\alpha+\beta+\gamma)p)-3p+2p\log 2p+p\log p$$

Thus, we must now solve the system of equations

$$\begin{cases} -2p\log(2\alpha p+\beta p)-p\log((\alpha+\beta+\gamma)p)-3p+2p\log 2p+p\log p=\lambda \\ -p\log(2\alpha p+\beta p)-p\log((\alpha+\beta+\gamma)p)-p\log(\beta p+2\gamma p)-3p-3p\log p=\lambda \\ -2p\log(\beta p+2\gamma p)-p\log((\alpha+\beta+\gamma)p)-3p+2p\log 2p+p\log p=\lambda \\ \alpha+\beta+\gamma=1 \end{cases}$$

resulting in the following solution.

The capacity of the channel in Figure 4-7 is given by $\Lambda=(1-p)\log\left(\frac{1-p}{2}\right)+(1-p)\log(1-p)$, **which corresponds to an input distribution of** $\{\alpha=1/2,\beta=0,\gamma=1/2\}$.

Hence, the capacity of the channel when $p=1/3$ is $\Lambda=2/3$ bits per symbol corresponding to an input distribution of $\{P(X=x)\}=\{1/2, 0, 1/2\}$, or alternatively, an output distribution of $\{P(Y=y)\}=\{1/3, 1/3, 1/3\}$. In physical terms, we can achieve the capacity of this channel if we do not use the second input, and use the other two inputs with equal probability.

We note here that, from Theorem 4.10, we could have actually assumed at the outset that $\alpha=\gamma$, thus simplifying our calculations as shown below.

Example 4.13 Revisited: Using Theorem 4.10, we have that $\alpha=\gamma$. Therefore, we can calculate

$$H(Y\mid X)=-2\alpha\left(\frac{2}{3}\log\left(\frac{2}{3}\right)+\frac{1}{3}\log\left(\frac{1}{3}\right)\right)-(1-2\alpha)\left(\log\left(\frac{1}{3}\right)\right)=-\frac{4\alpha}{3}+\log 3$$

Upon calculating the output probabilities in order to determine $H(Y)$, we see that

$$P(Y=0)=\frac{2}{3}\alpha+\frac{1}{3}(1-2\alpha)=\frac{1}{3}=P(Y=1)=P(Y=2)$$

Therefore, $H(Y) = \log 3$, so that $I(Y;X) = \log 3 + \frac{4\alpha}{3} - \log 3$. Hence, capacity will be achieved when α is as large as possible. In this case, since we have both $\alpha \geq 0$ and $1-2\alpha \geq 0$, then the resulting capacity of the channel is $\Lambda = 2/3$ corresponding to an input distribution of $\{P(X = x)\} = \{1/2, 0, 1/2\}$ as above.

This particular channel appears in Goldie & Pinch (pg. 120, [8]), and while the authors are indeed correct in asserting that capacity is achieved by setting one of the outputs to '0', they have made a mistake in calculating the actual capacity of the channel, which they assert incorrectly to be $\Lambda = \log 3$.

Note that this would only be possible if either the inputs or outputs were uniformly distributed *and* there was no noise present in either the channel from $X \rightarrow Y$ or $Y \rightarrow X$ which is clearly not the case.

Another way of showing that the capacity calculated by Goldie & Pinch is incorrect is to consider the fact that the only way to achieve a capacity of $\Lambda = \log 3$ is to have the case where each input gets mapped to a distinct output with probability 1, as shown below in Figure 4-8.

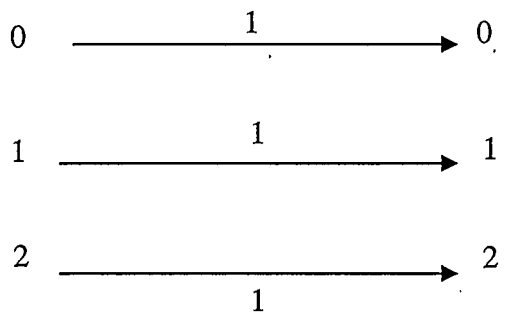

Figure 4-8 An example of a noiseless channel

Analytically speaking, this amounts to the following theorem.

Theorem 4.14 (Capacity of a Noiseless Channel): *Given a channel* $\Gamma : \mathrm{X} \rightarrow \mathrm{Y}$, *where the inputs are given by* $X = \{x_1, x_2, ..., x_n\}$ *and the outputs are given by* $Y = \{y_1, y_2, ..., y_n\}$, *a capacity of* $\Lambda = \log n$ *can be obtained if and only if* $H(Y \mid X) = 0$ *and the outputs are uniformly distributed.*

Note that Theorem 4.14 corresponds to a noiseless channel, where each of the *n* output fans are disjoint.

Example 4.15: Now consider the ternary channel shown in Figure 4-9, that also appears in Goldie & Pinch (pg .120, [8]), albeit with an incorrect conclusion.

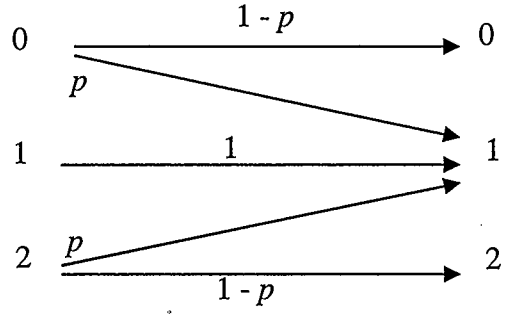

Figure 4-9 Ternary channel with one deterministic input

The corresponding channel transition matrix in this case is then

$$P := \begin{bmatrix} 1-p & p & 0 \\ 0 & 1 & 0 \\ 0 & p & 1-p \end{bmatrix}$$

Again, we can proceed using the method of Lagrange Multipliers, and since the channel is not symmetric, this method is our best bet. In this case, using an input distribution of $\{P(X = x)\} = \{\alpha, \beta, \gamma = 1 - \alpha - \beta\}$, we can compute

$$P(Y=0)=\sum_{x\in X}P(Y=0\mid X=x)P(X=x)=P(Y=0\mid X=0)P(X=0)+\ldots+P(Y=0\mid X=2)P(X=1)$$

$$=\alpha(1-p)$$

Similarly, we can compute $P(Y=1)=\alpha p+\beta+\gamma p, P(Y=2)=\gamma(1-p)$. Hence, we have

$$H(Y)=-(\alpha-\alpha p)\log(\alpha-\alpha p)-(\alpha p+\beta+\gamma p)\log(\alpha p+\beta+\gamma p)-(\gamma-\gamma p)\log(\gamma-\gamma p)$$

Again, using the entries of the channel transition matrix, we calculate

$$H(Y\mid X)=-P(X=0)P(Y=0\mid X=0)\log P(Y=0\mid X=0)-\ldots-P(X=2)P(Y=2\mid X=2)\log P(Y=2\mid X=2)$$

$$=-\alpha\big((1-p)\log(1-p)+p\log p\big)-\beta(1)\log(1)-\gamma\big(p\log p+(1-p)\log(1-p)\big)$$

$=(\alpha+\gamma)H(p)$ where $H(p)$ is the Shannon function.

Then, we can compute the capacity of the channel by using the method of Lagrange Multipliers, with objective function $f(\alpha,\beta,\gamma)=H(Y)-H(Y\mid X)$ and constraint function given by $g(\alpha,\beta,\gamma)=\alpha+\beta+\gamma-1$. Computing the partial derivatives of f, we have

$$f_\alpha(\alpha,\beta,\gamma)=-(1-p)\log(\alpha-\alpha p)-p\log(\alpha p+\beta+\gamma p)-p-H(p)$$
$$f_\beta(\alpha,\beta,\gamma)=-\log(\alpha p+\beta+\gamma p)-1$$
$$f_\gamma(\alpha,\beta,\gamma)=-(1-p)\log(\gamma-\gamma p)-p\log(\alpha p+\beta+\gamma p)-p-H(p)$$

Then, by solving the system of equations

$$\begin{cases}-(1-p)\log(\alpha-\alpha p)-p\log(\alpha p+\beta+\gamma p)-p-H(p)=\lambda\\ -\log(\alpha p+\beta+\gamma p)-1=\lambda\\ -(1-p)\log(\gamma-\gamma p)-p\log(\alpha p+\beta+\gamma p)-p-H(p)=\lambda\\ \alpha+\beta+\gamma=1\end{cases}$$

we obtain a capacity of $\Lambda = (1-p)\log\left(\frac{1-p}{2}\right) + (1-p)\log(1-p)$, which is achieved with an input distribution of $\{P(X = x)\} = \{1/3, 1/3, 1/3\}$ or equivalently, an output distribution of $\{P(Y = y)\} = \{1/3, 1/3, 1/3\}$.

In their book, Goldie & Pinch claim that, for the above channel, a capacity of $\Lambda = \log 3$ can be achieved for any value of $p < 1/3$. This statement is, for the most part erroneous, with the exception of the case when $p = 0$ for which the capacity is indeed $\Lambda = \log 3$.

Note that using Theorem 4.14, we can compute a capacity of $\Lambda = \log 3$ exactly when the outputs are uniformly distributed and $p = 0$, since this choice of p would result in a noiseless channel, i.e. $H(Y \mid X) = 0$. We also point out here that using Theorem 4.10, we could have assumed at the outset that $\alpha = \gamma$, thereby reducing a three-variable optimization problem to a one-variable problem as was done in Example 4.13.

As we have seen, the calculation of capacity for certain channels can be an elusive task, especially if one makes one or more misguided assumptions along the way. We point out here that the use of Lagrange Multipliers, or its more generalized version, the Kuhn-Tucker algorithm, is the most common method used to evaluate the capacity of discrete channels. However, both the Regular Channel Theorem and the Near-Regular Channel Theorem can be used to exploit symmetry in a given channel to greatly reduce the amount of computation required to obtain the capacity of the channel. More specifically, the Near-Regular Channel Theorem allows us to reduce a capacity problem involving n variables to one involving only one variable. We have also shown here that in some cases, there may exist infinitely many input distributions that can be used to achieve capacity. This contradicts the implication by some that in order to achieve the capacity of a channel, one must have a unique set of input probabilities.

5 Reciprocal Channels

First we pose the following problem. We are given a binary channel Γ' with input probability $p_0 = P(X=0)$ and $q_0 = 1-p_0 = P(X=1)$. Let p be a constant with $0 < p < 1/2$. We assume that p_0 satisfies the constraint

$$p < p_0 < 1-p \tag{5.1}$$

It then follows that

$$p < q_0 < 1-p \tag{5.2}$$

Suppose the channel matrix P is given by

$$\mathrm{P} = \begin{bmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{bmatrix}$$

where each row of P adds up to 1. Moreover, we assume that

$$P_{00} = \frac{(1-p)(p_0-p)}{p_0(1-2p)} \text{ and } P_{11} = \frac{(1-p)(q_0-p)}{q_0(1-2p)}$$

We note that the channel matrix P depends on the input probabilities: it is not a constant as is usually the case. Then, with $\alpha = \frac{p_0-p}{1-2p}$ and $\beta = 1-\alpha = \frac{q_0-p}{1-2p}$, we have

$$P = \begin{bmatrix} \dfrac{(1-p)\alpha}{(1-p)\alpha+p\beta} & \dfrac{p\alpha}{p\alpha+(1-p)\beta} \\ \dfrac{p\beta}{(1-p)\alpha+p\beta} & \dfrac{(1-p)\beta}{p\alpha+(1-p)\beta} \end{bmatrix}$$

It must first be noted that this channel matrix is **not** in general symmetric (with the exception of the case when $\alpha = \beta$), so we cannot make use of Theorem 4.7 to determine the capacity.

Hence, we can proceed by way of two methods, either by using Ash's formula for general binary channels, or by using Lagrange Multipliers.

Recall that via Theorem 4.6, the capacity of a general binary channel can be computed as

$$\Lambda = \log\left(2^{u} + 2^{v}\right)$$

Unfortunately, the actual expression for the capacity is quite cumbersome and depends on two variables, particularly p and α. Hence, we can **not** obtain a closed form for the capacity unless the input probabilities specified first. The relevant calculations are included in Appendix A.

In an attempt to find a more tidy solution, we can try the method of Lagrange Multipliers, as was done for the channels in Examples 4.13 and 4.15.

To compute the capacity of this channel, we find $H(Y)$ by regarding Y as a source as before:

$$P(Y=0) = \sum_{x \in X} P(Y=0 \mid X=x)P(X=x) = P(Y=0 \mid X=0)P(X=0) + P(Y=0 \mid X=1)P(X=1)$$

$$= \frac{(1-p)\alpha}{(1-p)\alpha + p\beta}((1-p)\alpha + p\beta) + \frac{p\alpha}{p\alpha + (1-p)\beta}(p\alpha + (1-p)\beta)$$

$$= (1-p)\alpha + p\alpha = \alpha$$

Similarly, we compute $P(Y=1) = P(Y=1 \mid X=0)P(X=0) + P(Y=1 \mid X=1)P(X=1) = \beta$. Thus we have $H(Y) = -\alpha \log \alpha - \beta \log \beta = H(\alpha)$.

Next, we must determine $H(Y \mid X)$ and find the maximum value of $I(Y;X) = H(Y) - H(Y \mid X)$.

By definition, we have that $H(Y|X) = \sum_{x \in X} P(X=x) \sum_{y \in Y} P(Y=y|X=x) \log P(Y=y|X=x)$.

Using the entries of P, we have

$$H(Y|X) = -P(X=0)P(Y=0|X=0)\log P(Y=0|X=0) - \ldots - P(X=1)P(Y=1|X=1)\log P(Y=1|X=1)$$

$$= -\left((1-p)\alpha + p\beta\right)\frac{(1-p)\alpha}{(1-p)\alpha + p\beta}\log\frac{(1-p)\alpha}{(1-p)\alpha + p\beta} - \ldots - \left(p\alpha + (1-p)\beta\right)\frac{(1-p)\beta}{p\alpha + (1-p)\beta}\log\frac{(1-p)\beta}{p\alpha + (1-p)\beta}$$

$$= -(1-p)\alpha\log\frac{(1-p)\alpha}{(1-p)\alpha + p\beta} - p\beta\log\frac{p\beta}{(1-p)\alpha + p\beta} - p\alpha\log\frac{p\alpha}{p\alpha + (1-p)\beta} - (1-p)\beta\log\frac{(1-p)\beta}{p\alpha + (1-p)\beta}$$

Using Maple, no closed form for capacity can be obtained (see Appendix A for corresponding calculations). The reason that Lagrange fails to provide a solution is that there are extra constraints involved here that were not present in the examples above from Goldie & Pinch, namely constraints (5.1) and (5.2) above. With our two most promising methods exhausted, what other option do we have?

Fortunately, with a bit of convenient manipulation, we can make use of the **symmetry property of mutual information** to solve the problem with great ease. As it turns out, the channel matrix given above corresponds to the induced, or reciprocal, channel Γ' (i.e. $\Gamma' : Y \to X$ instead of $\Gamma : X \to Y$) of a binary symmetric channel Γ with parameter p and input probabilities $P(X=0) = \alpha$, $P(X=1) = \beta = 1-\alpha$. Note that constraints (5.1) and (5.2) above fall out here.

As we have already seen, the capacity of the BSC is given by $\Lambda = 1 - H(p)$. Thus, because of the crucial fact that $\Lambda = \max_{P(X=x)} I(X;Y) = \max_{P(Y=y)} I(Y;X)$ (as discussed in Chapter 3 and 4), we can compute the capacity of the channel Γ' above to be

$$\Lambda = \max_{P(X=x)} I(X;Y) = \max_{P(Y=y)} I(Y;X) = \max_{P(Y=y)} \left(H(Y) - H(Y|X)\right) = 1 - H(p) = capacity(\Gamma')$$

and we have completely solved the problem!

It is important to emphasize, as above, that the reciprocal of a BSC is **not** a BSC, with the exception of the case when $\alpha = \beta$.

In general, an induced or reciprocal channel is a channel $\Gamma' : Y \to X$ with transition probabilities given by $P(X = x \mid Y = y)$ for all input values x and output values y corresponding to a channel $\Gamma : X \to Y$. In the example above, the input probabilities of the induced channel were obtained using Bayes' Formula from the input probabilities of the original binary symmetric channel Γ and the entries of the binary symmetric channel transition matrix

$$P = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}$$

Note that using Bayes' Formula, it can be shown that the reciprocal of the reciprocal of a binary symmetric channel Γ is Γ.

Theorem 5.1 (Reciprocal of the Reciprocal of a BSC): *Given a BSC with* $P(X = 0) = \alpha$, $P(X = 1) = \beta = 1 - \alpha$, *then the reciprocal of the reciprocal of a BSC is a BSC.*

Proof: We have already seen that a reciprocal BSC defined by $\Gamma' : Y \to X$ with input probabilities given by $P(Y = 0) = (1-p)\alpha + p\beta$, $P(Y = 1) = p\alpha + (1-p)\beta$, where $\beta = 1 - \alpha$ has a channel transition matrix given by

$$P = \begin{bmatrix} \dfrac{(1-p)\alpha}{(1-p)\alpha + p\beta} & \dfrac{p\alpha}{p\alpha + \beta(1-p)} \\ \dfrac{p\beta}{(1-p)\alpha + p\beta} & \dfrac{(1-p)\beta}{p\alpha + (1-p)\beta} \end{bmatrix}$$

Therefore, by viewing the output of the reciprocal channel as a source, we have

$$P(X = 0) = P(X = 0 \mid Y = 0)P(Y = 0) + P(X = 0 \mid Y = 1)P(Y = 1)$$

$$= \frac{(1-p)\alpha}{(1-p)\alpha + p\beta}((1-p)\alpha + p\beta) + \frac{p\alpha}{p\alpha + (1-p)\beta}(p\alpha + (1-p)\beta) = \alpha$$

$$P(X=1) = P(X=1 \mid Y=0)P(Y=0) + P(X=1 \mid Y=1)P(Y=1)$$

$$= \frac{p\beta}{(1-p)\alpha + p\beta}((1-p)\alpha + p\beta) + \frac{(1-p)\beta}{p\alpha + (1-p)\beta}(p\alpha + (1-p)\beta) = \beta$$

Also, using Bayes' formula, the conditional probabilities $P(Y=y \mid X=x)$ can be calculated as

$$P(Y=0 \mid X=0) = \frac{P(X=0 \mid Y=0)P(Y=0)}{P(X=0)} = \frac{(1-p)\alpha(1-p)\alpha + p\beta}{((1-p)\alpha + p\beta)\alpha} = 1-p$$

Similarly, $P(Y=1 \mid X=0) = p$, $P(Y=0 \mid X=1) = p$, and $P(Y=1 \mid X=1) = 1-p$.

Therefore, upon flipping the reciprocal BSC, the input probabilities, coupled with the conditional probabilities $P(Y=y \mid X=x)$, correspond exactly to those of a BSC, hence the underlying structure of a channel is not changed when it is flipped.

□

In connection with the above result, we have the following more general theorem.

Theorem 5.2 (Reciprocal of the Reciprocal of a Discrete Channel): *Given a discrete channel* $\Gamma : X \to Y$, *the reciprocal of the reciprocal of* Γ *is* Γ.

Sketch of Proof: Given a channel Γ, we have the channel transition probabilities $P(Y=y \mid X=x)$ for all $x \in X, y \in Y$. In the reciprocal channel Γ', the channel transition probabilities are

$$P(X=x \mid Y=y) = \frac{P(Y=y \mid X=x)P(X=x)}{P(Y=y)} \qquad (5.3)$$

Now, when we reciprocate Γ', we obtain the transition probabilities

$$\frac{P(X = x \mid Y = y)P(Y = y)}{P(X = x)} = \frac{P(Y = y \mid X = x)P(X = x)P(Y = y)}{P(Y = y)P(X = x)} = P(Y = y \mid X = x)$$

which are exactly the transition probabilities corresponding to the channel Γ. Hence, by reciprocating the reciprocal of Γ, we obtain Γ.

□

Corollary 5.3: *Every channel is the reciprocal of a (unique) channel.*

Proof: Given a channel $\Gamma : X \rightarrow Y$, from Theorem 5.2 we have that $\Gamma = \text{reciprocal}(\Gamma')$. To show that Γ is indeed unique, suppose that $\Gamma = \text{reciprocal}(M)$ and $\Gamma = \text{reciprocal}(N)$ where M, N are distinct channels. Then, by reciprocating both sides of each expression, we obtain

$$\text{reciprocal}(\Gamma) = M \text{ and reciprocal}(\Gamma) = N$$

which implies that $M = N$. But, we assumed that M, N were distinct. Therefore, by contradiction, we have our result.

□

Next, we discuss a practical example. Consider the use of diagnostic equipment that is used to detect underlying input conditions based on detected output signals. As discussed above, it is possible to regard the output of a channel as a source by using the Law of Total Probability. Then, using Bayes' Formula, it is possible to calculate the transition probabilities of the induced channel, since for each input x and output y, we have

$$P(X = x \mid Y = y) = \frac{P(Y = y \mid X = x)P(X = x)}{P(Y = y)}$$

Thus, by viewing the results of a diagnostic test, we are actually making use of an induced channel in the sense that viewing the output of the original channel gives us much insight into the input of the original channel.

To appreciate the usefulness of induced channels in a practical setting, consider the following example from Luenberger [14].

***Example 5.4*:** Consider an oil company that has discovered a potentially promising site to drill for oil. At this site, there are two possibilities: either oil is present at the site with a probability of 1/3, or the well is "dry" and no oil is present with a probability of 2/3. Assume that if the well is "wet", the company stands to gain \$600 million in oil reserves. Alternatively, a dry well results in \$0 return. Also suppose it costs \$120 million to drill a well.

Before evaluating channel information, the expected payoff to the oil company is given by

$$\begin{aligned} E(payout) &= P(wet)(\$600M) + P(dry)(\$0) - \text{cost of drilling} \\ &= (1/3)(\$600M) + (2/3)(\$0) - \$120M = \$80M \end{aligned}$$

Now, suppose the most technologically advanced seismic imaging technology has revealed that there is a strong possibility that oil is present at the site. In particular, if a positive test result is observed, there is a 75% chance that oil is present, and if a negative test result is observed, there is a 75% chance that the site is dry. With *A, B* representing the output and input respectively of the channel, the corresponding channel matrix is then

$$P(B \mid A) = \begin{array}{c} \\ wet \\ dry \end{array} \begin{array}{c} \begin{array}{cc} pos. & neg. \end{array} \\ \begin{bmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{bmatrix} \end{array}$$

The decision to drill is based on the outcome of the seismic testing, which unfortunately in this example, is not as accurate as oil executives would hope. In actual fact, the probability of a profitable decision can be greatly increased if we make use of the induced channel.

Regarding the test outcomes as a source, we have

$$P(positive) = P(positive \mid wet)P(wet) + P(positive \mid dry)P(dry)$$
$$= (3/4)(1/3) + (1/4)(2/3) = 5/12$$

$$P(negative) = P(negative \mid wet)P(wet) + P(negative \mid dry)P(dry) = 1 - P(positve) = 7/12$$

The channel transition probabilities of the induced channel are calculated using Bayes' Formula:

$$P(wet \mid positve) = \frac{P(positive \mid oil)P(oil)}{P(positive)} = \frac{(3/4)(1/3)}{(5/12)} = \frac{3}{5}$$

Similarly, we can compute

$$P(wet \mid negative) = \frac{1}{7},\ P(dry \mid positive) = \frac{2}{5},\ \text{and}\ P(dry \mid negative) = \frac{6}{7}.$$

Now, supposing it costs roughly $120 million to drill for oil, what decision should the company make based on the test results?

If the test result is positive, we see that there is a 3 in 5 chance that oil is actually present. Thus the expected value of the payoff is (3/5)($600 million)-($120 million) = $240 million. On the other hand, if the test is negative, there is a 1 in 7 chance that oil is actually present. The expected return in this case is (1/7)($600 million)-($120 million) = -$34.29 million. Overall, the expected net profit is calculated to be $100 million, which is $20 million dollars greater than the expected profit obtained before making use of the induced channel. Thus, it is clear that, based on the probabilities of false negatives and false positives, if the test is

positive, the company should drill, and if it is negative, the company should stop. Also note that, in each case, the actual act of testing likely costs the oil company in the form of consulting fees, although compared to the amount of the potential payout, the cost can be considered negligible.

Thus, using the induced channel, the oil company can decide whether to drill or not based on the results of the seismic test, as opposed to blindly drilling based on the *a priori* probabilities of oil being present at the site in question. While this may seem to be an obvious choice, this example serves to exhibit the value of induced channels and the information that they convey.

6 Signal Processing

Another fundamental aspect of Information Theory pertains to signal processing, the technique of transmitting and receiving continuous streams of data. The theoretical basis for signal processing is centred around the celebrated sampling theorem. In this chapter, the definition of entropy for continuous random variables will be presented, along with an analytical sketch of the proof of Shannon's famous capacity result. We will then move on to provide a brief discussion of the sampling theorem, coupled with a generalization of the theorem to include signals sampled in a practical manner. The main purpose of this chapter is to provide us with some context in order to investigate a novel cryptosystem due to Alan Turing in Chapter 7.

6.1 Shannon's Capacity Theorem for Continuous Channels

We now revisit the concept of channel capacity, this time from a continuous point of view. Recall that in the discrete case, the capacity was essentially the maximum value of the entropy of the signal *and* the noise, or $H(Y)$, minus the conditional entropy of the noise itself, or $H(Y \mid X)$. The same argument holds true for continuous channels as well, although we first require a definition for continuous entropy in order to proceed.

Definition 6.1: *The* ***entropy*** *of a continuous random variable X with probability density* $p(x)$ *is defined as*

$$h(x) = \int_{-\infty}^{\infty} p(x) \log p(x) dx \qquad (6.1)$$

The subtlety involved with this definition is that as shown, it only represents part of the entropy one would obtain if the continuous signal was discretized, where values of the function are taken at intervals of Δx with $\Delta x \to 0$. The other part of the "discretized entropy" is infinite, but it doesn't depend on probabilities. Furthermore, when we apply the definition

above, it will be in the form of a difference of entropies, and the infinite part of each entropy in the difference will cancel out. Thus, by simply ignoring the infinite part, the above definition is appropriate.

Also, recall that in Chapter 2, the probability density for a Gaussian random variable with a mean of zero and variance σ^2 is given by

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-x^2}{2\sigma^2}\right) \tag{6.2}$$

It was also mentioned in Chapter 2 that random perturbations due to channel noise are best modeled using a Gaussian distribution with a mean of 0, and that the noise is additive and independent. Such noise is referred to as Additive White Gaussian Noise, or AWGN. Therefore, we can calculate the entropy of continuous channel noise as

$$\begin{aligned}
h(x) &= -\int_{-\infty}^{\infty} p(x) \log\left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-x^2}{2\sigma^2}\right)\right] dx \\
&= -\int_{-\infty}^{\infty} p(x) \log(e) \ln\left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-x^2}{2\sigma^2}\right)\right] dx \qquad (\text{since } \log x = \log(e)\ln x) \\
&= -\log(e) \int_{-\infty}^{\infty} p(x) \left(\ln\left[\frac{1}{\sqrt{2\pi\sigma^2}}\right] - \ln\left[\exp\left(\frac{-x^2}{2\sigma^2}\right)\right]\right) dx \\
&= -\log(e) \int p(x) \left(\frac{-x^2}{2\sigma^2} - \ln\sqrt{2\pi\sigma^2}\right) dx
\end{aligned}$$

Now, from [14], $\int_{-\infty}^{\infty} x^2 p(x) = \sigma^2$ and since $\int_{-\infty}^{\infty} p(x)dx = 1$, we can write

$$h(x) = \log(e)\left[\frac{1}{2} + \ln\sqrt{2\pi\sigma^2}\right] = \log(e)\left[\frac{1}{2}\ln e + \frac{1}{2}\ln 2\pi\sigma^2\right] = \frac{1}{2}\log(e)\ln 2\pi e\sigma^2 = \frac{1}{2}\ln 2\pi e\sigma^2$$

Before moving on, we will need to outline a property for Gaussian density from [14], by way of the following theorem.

Theorem 6.2 (Maximal Property of Gaussian Density): *The continuous entropy of a random variable, $h(x)$, subject to the constraints*

$$\int_{-\infty}^{\infty} p(x)dx = 1 \text{ and } \int_{-\infty}^{\infty} x^2 p(x)dx \leq S \tag{6.3}$$

where S is the average power of the signal, is maximized when $p(x)$ is Gaussian and $Var\{h(x)\} = \sigma^2 = S$.

Now, suppose we have a signal to be transmitted over a continuous, noisy channel subject to AWGN. Then with Y representing the received signal and Z the AWGN, then clearly we have $Y = X + Z$, where X and Z are assumed to be independent. This situation is depicted in Figure 6-1.

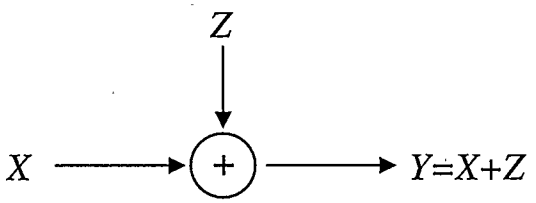

Figure 6-1 Signal transmission over a continuous channel subject to AWGN

Furthermore, we know that the capacity of the channel between X and Y is given by $\Lambda = \max I(X;Y) = \max I(Y;X) = \max\big(H(Y) - H(Y \mid X)\big)$ where this time, we are maximizing subject to the constraints outlined in Theorem 6.2. Thus, we arrive at one of Shannon's most famous theorems.

Theorem 6.3 (Continuous Channel Capacity): *Given a transmitted signal X with average power S, a received signal Y, and AWGN represented by Z such that Z has average power N, the capacity Λ of the channel $\Gamma : X \rightarrow Y$ subject to Z is given by*

$$\Lambda = \frac{1}{2}\log\left(1 + \frac{S}{N}\right) \tag{6.4}$$

and is achieved when the probability density of X, $p(x)$, is Gaussian.

Sketch of Proof: Given a transmitted signal X with average power S, a received signal Y, and AWGN represented by Z, where Z has average power N, in order to find the capacity, we must maximize

$I(X;Y)$ subject to the constraint $\int_{-\infty}^{\infty} p(x)dx = 1$, where $p(x)$ is the probability density of X.

Then, we can find the capacity of the channel by computing

$$\begin{aligned} I(X;Y) = H(Y) - H(Y \mid X) &= H(Y) - H(X+Z \mid X) \\ &= H(Y) - H(Z \mid X) \\ &= H(Y) - H(Z) \qquad \text{(since } X \text{ and } Z \text{ are independent)} \end{aligned}$$

and maximizing $I(X;Y) = H(Y) - H(Z)$ subject to $\int_{-\infty}^{\infty} p(x)dx = 1$ and $\int_{-\infty}^{\infty} x^2 p(x)dx \le S$.

From above, we know that $H(Z) = h(z) = \frac{1}{2}\log 2\pi e\sigma^2$ bits, and from Theorem 6.2, we know that $H(Y)$ is maximized when the signal has a Gaussian density with average power S.

Therefore, the capacity can be computed as

$$\Lambda = \max I(X;Y) = \max\big(H(Y) - H(Z)\big) = \max\big(H(Y)\big) - \frac{1}{2}\log 2\pi eN$$

Finally, since the noise is additive by definition, and X has a Gaussian density with average power S, we have

$$\Lambda = \max\big(H(Y)\big) - \frac{1}{2}\log 2\pi eN = \frac{1}{2}\log[2\pi e(S+N)] - \frac{1}{2}\log[2\pi e(N)] = \frac{1}{2}\log\left(1 + \frac{S}{N}\right)$$

□

With the continuous framework for signals and channels, we can now move on to some fundamental results pertaining to signal processing in general.

6.2 The Sampling Theorem

Consider a continuous, analog signal that one might wish to transmit from one point to another. For example, consider a temperature sensor mounted on a tower which measures the ambient temperature at a given site. Since temperature is measured in terms of real numbers, the resulting data stream of the sensor would be a continuous function, with an infinite and uncountable number of data points. Clearly, this data stream is of no practical use, since an infinite set of values is impossible to transmit! In order to obtain the desired temperature information, some estimation needs to be done, but the question remains, how can we estimate the stream of data to ensure that we have obtained all of the important data measured by the sensor using only a finite amount of information? The answer can be found in the celebrated sampling theorem, initially published by Claude Shannon, which acts as one of the cornerstones of communication engineering and is responsible for setting the groundwork for analog communications as we know it.

Through the use of Fourier transforms Shannon was able to determine, by building upon the work of Hartley and Nyquist, the exact number of data points, or "samples", that are required in order to reconstruct an analog signal using only these points. Using this theorem, a time-varying signal that is band-limited (i.e. it has no frequency components beyond a finite range) can be sampled at multiples of a basic sampling interval, and reconstructed upon transmission using only these sampled values and a reconstruction formula proposed by Shannon and Nyquist before him. This in itself is a fascinating result: a continuous stream of data with an infinite and uncountable amount of information can be accurately represented using only a countable set of data points. For a fully detailed proof of the sampling theorem, refer to [17] or [13].

Theorem 6.4 (The Sampling Theorem): *A band-limited signal $x(t)$ of finite energy, which has no frequency components higher than W Hertz, is completely described by specifying the values of the signal at instants of time separated by $1/2W$ seconds. Furthermore, it is possible to reconstruct the original signal from the specified values of $x(t)$ using the reconstruction formula*

$$x(t) = \sum_{n=-\infty}^{\infty} 2W_1 T_s x(nT_s)\text{sinc}[2W_1(t - nT_s)]$$

where W_1 is an arbitrary number such that $W \le W_1 \le \frac{1}{T_s} - W = f_s - W$, and $T_s = 1/f_s$ is the sampling interval with $T_s \le 1/2W$.

Sketch of Proof: Assume that $x(t)$, a signal of finite energy, satisfies the Dirichlet conditions outlined in Chapter 2, and suppose that $x(t)$ is an arbitrary band-limited signal with bandwidth W and amplitude A as shown below in Figure 6-2.

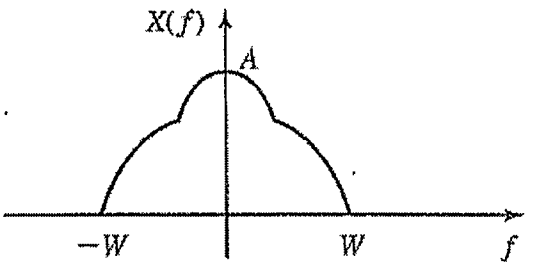

Figure 6-2 Frequency domain representation of an arbitrary signal $x(t)$ ([17])

Let $x_\delta(t)$ be the result of the sampling process upon sampling $x(t)$ at nT_s time instants, where n is an integer. Then we have

$$x_\delta(t) = \sum_{n=-\infty}^{\infty} x(nT_s)\delta(t - nT_s) \text{ where } \delta(t - nT_s) = \begin{cases} 1, \text{ if } t = nT_s \\ 0, \text{ if } t \neq nT_s \end{cases}$$

Thus, we can write $x_\delta(t) = x(t)\sum_{n=-\infty}^{\infty}\delta(t-nT_s)$ without loss of generality.

Upon taking the Fourier transform of both sides and using the convention that $X(f)=\Im\{x(t)\}$ and $X_\delta(f)=\Im\{x_\delta(t)\}$, we obtain

$X_\delta(f) = X(f) * \Im\left\{\sum_{n=-\infty}^{\infty}\delta(t-nT_s)\right\}$ using the Convolution Property for Fourier Transforms.

Also, since $\Im\left\{\sum_{n=-\infty}^{\infty}\delta(t-nT_s)\right\}=\frac{1}{T_s}\sum_{n=-\infty}^{\infty}\delta(f-\frac{n}{T_s})$ from [17], we have

$$X_\delta(f)=X(f)*\frac{1}{T_s}\sum_{n=-\infty}^{\infty}\delta(f-\frac{n}{T_s}) \tag{6.5}$$

Using the convolution property of the impulse function, which states that $x(t)*\delta(t-t_0)=x(t-t_0)$ (see [14] for details), we get

$$X_\delta(f)=\frac{1}{T_s}\sum_{n=-\infty}^{\infty}X(f-\frac{n}{T_s}) \tag{6.6}$$

which represents a series of waveforms in the frequency domain of bandwidth W, centred about the frequencies $f = n/T_s$ for all integers n as shown in Figure 6-3.

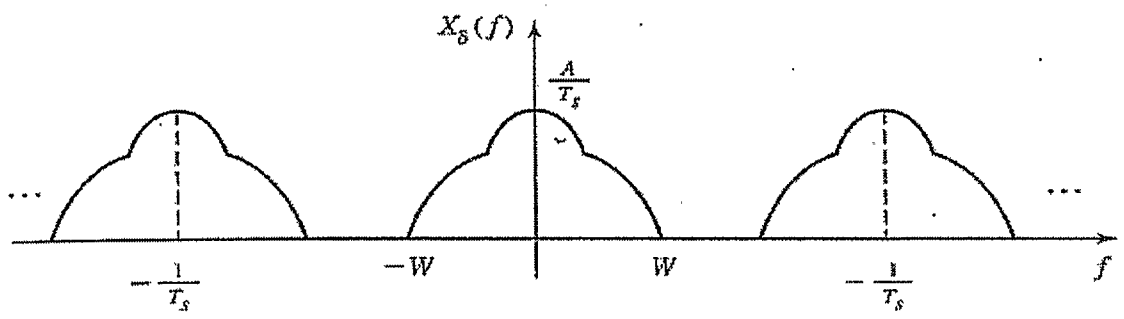

Figure 6-3 Frequency representation of the sampled signal corresponding to $x(t)$ ([17])

Note that there is a gap between each of the individual waveforms. This gap is referred to as the *guard band* of the signal, where the guard band is of width $f_s - 2W$ and is present as long as the samples are taken according to the *Nyquist criterion* $T_s \leq 1/2W$.

However, if the samples are taken at a rate *less* than the Nyquist rate, the signal waveforms will overlap and no amount of filtering will allow us to recover the original spectrum $X(f)$. This over-lapping error is referred to as *aliasing error*, and can be prevented only if a signal is sampled such that at least two samples occur in one period. This condition is exactly satisfied by sampling according to the Nyquist criterion. For a more detailed account of aliasing, refer to Hamming ([9]).

On the other hand, samples taken at any rate *greater* than the Nyquist rate will permit unambiguous detection of the signal. This is analogous to the concept of input/output fans discussed in Chapter 4, where a given signal could be unambiguously detected as long as the input or output fans corresponding to the signal were disjoint.

Remark: The assertion that $x(t)$ is of finite energy rules out more troublesome functions such as $x(t) = \sin(t)$ or $x(t) = \cos(t)$. The assumption can be relaxed to allow for such functions if we sample at a rate greater than the Nyquist rate. For more details, see [13].

We continue with the sketch proof of Theorem 6.4. In order to extract the component that is centred about the origin (which is a scaled version of $X(f)$), we need to employ an ideal low-pass filter which "passes" all frequency components in a signal that are lower than the parameter, W_1, of the filter, and negates all others.

Mathematically speaking, in order to filter a signal with such a device, we multiply the transform of the signal with the transform of the filter, shown in Figure 6-4, along with the inverse transform of the filter in Figure 6-5.

$$H(f)=\begin{cases}T_s, |f|<W_1, with\ W\le W_1<f_s-W\\ 0, |f|\ge f_s-W\end{cases}$$

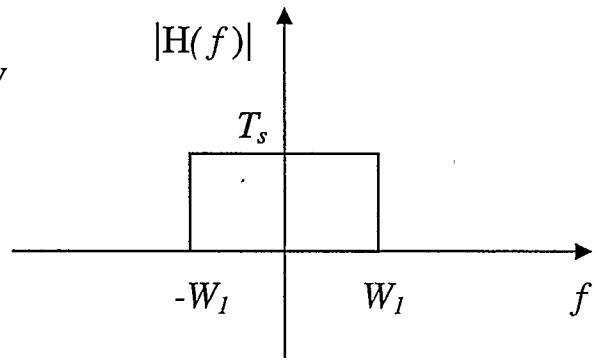

Figure 6-4 An ideal low pass filter

$$f(t)=\text{sinc}(t)=\frac{\sin(\pi t)}{(\pi t)}$$

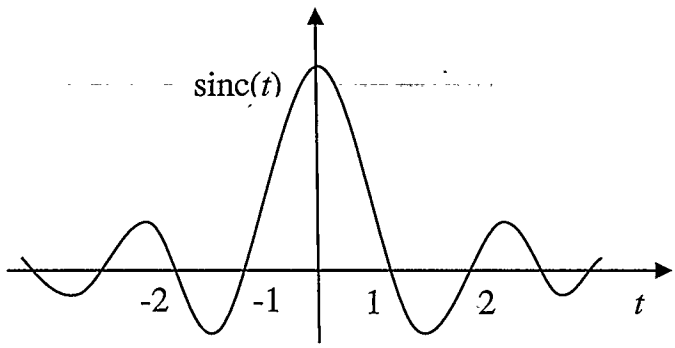

Figure 6-5 Graphical representation of the sinc function

Hence, with $X(f)=X_\delta(f)H(f)$, and from [17], $\Im^{-1}\{H(f)\}=2W_1T_s\text{sinc}(2W_1t)$, then taking the inverse transform of both sides yields

$$x(t)=x_\delta(t)*2W_1T_s\text{sinc}(2W_1t)=\sum_{n=-\infty}^{\infty}x(nT_s)\delta(t-nT_s)*2W_1T_s\text{sinc}(2W_1t)$$

Again, using the convolution property of the impulse function, we obtain the reconstruction formula

$$x(t)=\sum_{n=-\infty}^{\infty}2W_1T_sx(nT_s)\text{sinc}(2W_1(t-nT_s)) \qquad (6.7)$$

□

It must be noted that there are, unfortunately, some repercussions involved with translating the sampling theorem from the ideal, abstract situation to that of practical applications. These considerations are addressed further in [17] and [9].

6.2.1 The Sampling Theorem & the Capacity of a Band-Limited Channel

According to the sampling theorem, we must sample a given signal at a sampling rate of $T_s \leq 1/2W$. Furthermore, we know that the capacity of a Gaussian channel with a single input sample is

$$\Lambda = \frac{1}{2}\log\left(1+\frac{S}{N}\right) \tag{6.8}$$

Since we can have, at most, $2W$ distinct samples as per the Nyquist criterion, the capacity of a continuous, band-limited channel with bandwidth W subject to AWGN noise is then

$$\Lambda = 2W \cdot \frac{1}{2}\log\left(1+\frac{S}{N}\right) = W\log\left(1+\frac{S}{N}\right)$$

We now have the following theorem.

Theorem 6.5 (Capacity of a Band-Limited AWGN Channel): *Given a continuous band-limited signal X of bandwidth W and average power S, a received signal Y, and AWGN represented by Z such that Z has average power N, the capacity* Λ *of the channel* $\Gamma : X \rightarrow Y$ *subject to Z is given by*

$$\Lambda = W\log\left(1+\frac{S}{N}\right) \tag{6.9}$$

6.3 A More General Approach to Sampling

The sampling theorem as stated in Section 6.2 is an extremely important result, although it is only a limiting case in a more general theory. We can extend the theory to include signals sampled in a practical sense as well, which for our intents and purposes, are signals that are mixed with an arbitrary pulse train so as to capture samples which are pulses of period T and amplitude corresponding to the original signal. This method represents a more practical realization of sampling, as Theorem 6.4 above pertains to sampling instantaneously, which is not realizable in practice.

Consider the series of pulses, or pulse train, in Figure 6-6, along with the signal $x(t)$ shown below in Figure 6-7. By mixing the two signals, the net effect is that the pulses stretch or shrink in terms of amplitude to match the amplitude of $x(t)$ for the duration of each pulse, as shown in Figure 6-8.

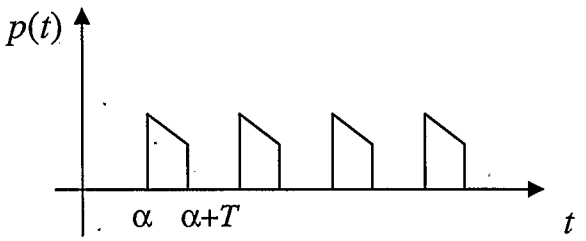

Figure 6-6 Arbitrary pulse train with period T

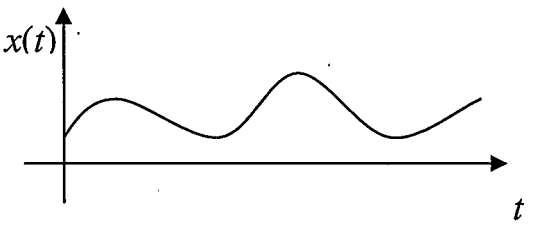

Figure 6-7 Time-varying, band-limited signal

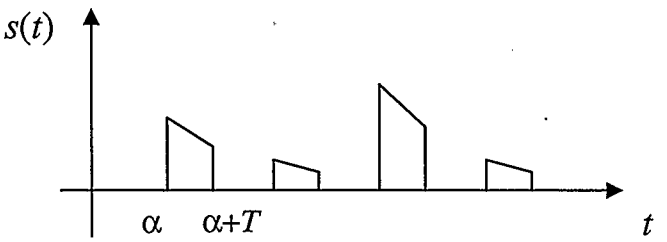

Figure 6-8 Mixed signal with period T

Analytically speaking, we have $s(t) = x(t)p(t)$, and it is assumed that $x(t)$ is a band-limited function with bandwidth W and is of finite energy, and that $p(t)$ is a periodic function of period T that satisfies the Dirichlet conditions outlined in Chapter 2.

Since $p(t)$ is periodic and satisfies the Dirichlet conditions, we can obtain the Fourier series expansion of $p(t)$, with $T = 1/f_p$:

$$p(t) = \sum_{n=-\infty}^{\infty} c_n e^{2\pi i n f_p t} \tag{6.10}$$

where $c_n = \frac{1}{T}\int_{\alpha}^{\alpha+T} p(t)e^{-2\pi i n f_p t}\,dt$ for some arbitrary α.

Then, we can compute the Fourier transform of $s(t)$ to be

$$S(f) = \int_{-\infty}^{\infty} x(t)\left[\sum_{n=-\infty}^{\infty} c_n e^{2\pi i n f_p t}\right] e^{-2\pi i f t}\,dt \tag{6.11}$$

Upon interchanging the summation and integration, we obtain

$$S(f) = \sum_{n=-\infty}^{\infty} c_n \int_{-\infty}^{\infty} x(t)e^{-2\pi i (f - n f_p) t}\,dt \tag{6.12}$$

Now, since $X(f) = \int_{-\infty}^{\infty} x(t)e^{-2\pi i f t}\,dt$, we can see that the integral in the expression for $S(f)$ is simply $X(f - nf_p)$, i.e. the frequency representation of $x(t)$ shifted by a frequency of nf_p for $n \in (-\infty, \infty)$.

Therefore,

$$S(f) = \sum_{n=-\infty}^{\infty} c_n X(f - nf_p) \tag{6.13}$$

which is essentially several copies of $X(f)$ centred at frequencies that are multiples of f_p as shown below in Figure 6-9, with amplitude determined by the complex Fourier coefficients c_n. As before, in order to prevent overlapping, we must choose f_p subject to the Nyquist criterion $f_p \geq 2W$. Then, we can capture the copy of $X(f)$ centred about the origin using a low-pass filter, and take the inverse Fourier transform to recover the signal $x(t)$ in a similar fashion as described in Section 6.2.

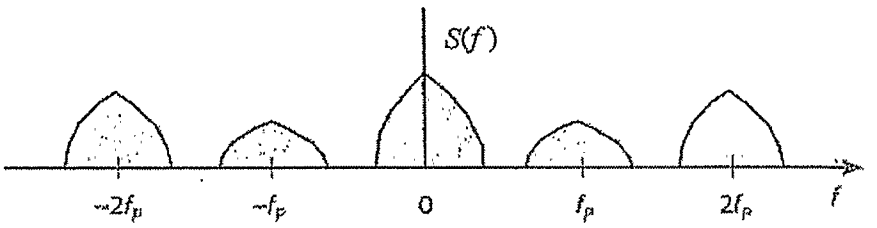

Figure 6-9 Fourier transform of the mixed signal $s(t)$ ([14])

A major practical consideration resulting the study of Fourier transforms is the fact that if a signal is band-limited (i.e. the Fourier transform of the signal is non-zero over a finite range of values), then it cannot be time-limited, and vice-versa. Unfortunately, since infinite-time and infinite-frequency signals cannot be used in a practical setting, the signals of interest must be truncated and approximated so that both the time-domain and frequency-domain signals are of finite support.

This problem comes up in the use of the sampling theorem, in that it is assumed that the time-varying signal is band-limited, implying that its time-domain representation has infinite support. However, this cannot occur in a practical sense. In order to rectify this apparent contradiction, the signals are subject to possibly significant amounts of error, due to the

widely-known Gibbs phenomenon for truncated signals (see Hamming ([9]) for more details). Hence, the sampling theorem and thus Shannon's capacity theorem for continuous channels, does indeed fit into the framework of theoretical signal processing, although practical considerations must be made in order to apply the sampling theorem to real-life applications.

With the survey of continuous signals now complete, we turn our focus to the application of Information Theory to the field of cryptography. We will soon see that by using Shannon's sampling theorem discussed here in conjunction with Shannon's idea for a perfectly secure discrete cryptosystem, that perfectly secure communications using analog signals is indeed possible, at least in principle.

7 Information Theory & Cryptography

Cryptography deals with the secure exchange of information between two parties, commonly referred to as Alice and Bob and denoted by **A** and **B** respectively. Seeking to intercept or disrupt this correspondence is an eavesdropper Eve, denoted by **E**, who is either a passive attacker (intercepts but does not interfere) or a malicious attacker (seeks to commit fraud by impersonation or otherwise). In classical cryptography, secure communication in a practical sense can be achieved using two main classes of cryptosystems, namely *symmetric* and *asymmetric* cryptosystems.

Asymmetric cryptosystems are based upon a publicly available key that is then used in conjunction with a mathematical problem that is assumed to be intractable such as the factorization of a very large number into a product of (typically two large) primes, to which only the recipient knows the answer. The security of such systems is based upon the assumed difficulty of the mathematical problem in question, and no rigorous proofs have been found to validate such assumptions, nor the security of an asymmetric cryptosystem. It must be noted that, although theoretically possible to break, public-key systems in use today are secure enough to be used in online banking, email, and countless other applications. The interested reader is referred to [2] for a more detailed overview of public-key cryptography.

Symmetric cryptosystems, on the other hand, are based on a private key possessed only by the sender and intended receiver(s) of a given message. In fact, it is then possible to design a cryptosystem which is totally unbreakable.

Our discussion regarding cryptography will focus on the examination of symmetric cryptosystems having **perfect secrecy**, a notion which ensures that, regardless of the computing power or the existence of a clever algorithm, a cryptosystem exhibiting such security will be unbreakable unless an eavesdropper possesses the key itself. A classical example of a cryptosystem offering perfect secrecy is the Vernam cipher, or *one-time pad*, which was developed by Claude Shannon. We will discuss the one-time pad before presenting a characterization of perfect secrecy using Latin squares under an assumption

concerning the cardinality of various sets. This characterization is not provided in the standard literature but is implicit in the work of Claude Shannon. Following this, we present new examples of perfect secrecy. At the end of the chapter, we briefly describe a rarely-discussed example of perfect secrecy for analog systems using Shannon's sampling theorem that goes back to Alan Turing.

7.1 Symmetric Cryptosystems and Perfect Secrecy

We begin our discussion here with the following definition.

Definition 7.1: *A symmetric cryptosystem* Γ *is a cipher system involving a finite set of possible messages* $M = \{m_1, m_2, ...\}$, *a finite set of encrypted messages, or ciphers* $C = \{c_1, c_2, ...\}$, *and a finite set of keyed enciphering transformations* $K = \{e_1, e_2, ...\}$, *where each* $e_i \in K$ *is a transformation that maps each* $m \in M$ *to each* $c \in C$ *injectively.*

We assume that each message has a non-zero probability of transmission, i.e. $P(M = m) > 0$ for each $m \in M$, otherwise we could simply delete it from our set of possible messages. Similarly, for each $c \in C$, it is assumed that there is at least one message that gets encrypted into c.

Furthermore, to decrypt, or undo the encryption, we apply the inverse of the enciphering transformation to a given cipher. Note that the inverse is well defined since each $e_i \in K$ is injective. Therefore, for any message m, we encrypt it as $e_i(m) = c$. To decrypt, we apply $d_i = e_i^{-1}$ to the cipher, i.e. $d_i(e_i(m)) = d_i(c) = e_i^{-1}(c) = m$.

Now, suppose we have a set of n possible messages, $M = \{m_1, m_2, ..., m_n\}$. Upon fixing the enciphering transformation e_k, we have that $\{e_k(m_1), e_k(m_2), ..., e_k(m_n)\}$ is a set of n distinct

ciphers. Therefore, the number of ciphers in C, denoted by $|C|$, is at least equal to the number of possible messages in M. Hence, we have that for a symmetric cryptosystem,

$$|M| \leq |C| \tag{7.1}$$

With the framework of a symmetric cryptosystem in hand, we can now discuss the desirable property of perfect secrecy for symmetric systems from an information-theoretic perspective. Perfect secrecy occurs when the ciphers produced by the cryptosystem tell us nothing new about the underlying message itself. Hence, if we view the encryption process as a discrete channel from our set of messages M to our set of ciphers C, we require that the capacity of the resulting channel is zero, or equivalently, the mutual information between the two random variables is zero. More formally, we have the following definition of perfect secrecy.

Definition 7.2 (Definition of Perfect Secrecy): *A symmetric cryptosystem* Γ *has perfect secrecy if*

$$I(M;C) = H(M) - H(M \mid C) = 0 \tag{7.2}$$

Thus for perfect secrecy we have $H(M) = H(M \mid C)$. From Chapter 2, we see that this is equivalent to saying that M and C are independent. So, we have the following equivalent definition of perfect secrecy.

Definition 7.3 (Equivalent Definition of Perfect Secrecy): *A symmetric cryptosystem* Γ *has perfect secrecy if M and C are independent, so that* $P(M = m \mid C = c) = P(M = m)$ *for all* $m \in M, c \in C$. *Alternatively, we write* $P(M \mid C) = P(M)$.

In what follows, for practical reasons, we use Definition 7.3. Let us now explore the implications of perfect secrecy. By definition, if we assume that our symmetric cryptosystem Γ exhibits perfect secrecy, then $P(M \mid C) = P(M)$ for all $m \in M, c \in C$. Now, let (m,c) be any message-cipher pair. Since $P(M = m) > 0$, then $P(M = m \mid C = c) > 0$. This implies that, with non-zero probability, the message m was encrypted into the cipher c.

Therefore, given any $m \in M, c \in C$, there exists *at least* one enciphering transformation e_k where $P(K = e_k) > 0$ such that $e_k(m) = c$.

Now, if we fix the message m and let the keys vary, then we have that the set $\{e_k(m) \mid e_k \in K\}$ contains all ciphers $c \in C$ and is thus equal to C, since each $e_k(m) \in C$. We can then conclude that

$$|C| \leq |K| \tag{7.3}$$

Furthermore, from (7.1), since $|M| \leq |C|$, we have that

$$|M| \leq |C| \leq |K| \tag{7.4}$$

Hence, to ensure that a symmetric cryptosystem exhibits perfect secrecy, we must have that the total number of enciphering transformations, or keys, be at least as big as the number of possible messages. We can now summarize with the following theorem.

Theorem 7.4 (Perfect Secrecy): *Given a symmetric cryptosystem* Γ *with messages* $M = \{m_1, m_2, ...\}$, *ciphers* $C = \{c_1, c_2, ...\}$, *and injective encryption keys* $K = \{e_1, e_2, ...\}$, *then a necessary condition for* Γ *to have perfect secrecy is* $|M| \leq |C| \leq |K|$.

We now turn to some examples of symmetric cryptosystems.

Example 7.5: Consider the symmetric cryptosystem Γ with messages $M = \{m_1, m_2, m_3, m_4\}$, ciphers $C = \{c_1, c_2, c_3, c_4\}$ and keys $K = \{e_1, e_2, e_3\}$ where the keys are defined as follows.

e_1:	e_2:	e_3:
$m_1 \to c_1$	$m_1 \to c_2$	$m_1 \to c_3$
$m_2 \to c_2$	$m_2 \to c_3$	$m_2 \to c_4$
$m_3 \to c_3$	$m_3 \to c_4$	$m_3 \to c_1$
$m_4 \to c_4$	$m_4 \to c_1$	$m_4 \to c_2$

Also, suppose the messages are distributed such that $P(M = m_1) = 0.1$, $P(M = m_2) = 0.2$, $P(M = m_3) = 0.3$, and $P(M = m_4) = 0.4$, and the keys are distributed such that $P(K = e_1) = 0.3$, $P(K = e_2) = 0.3$, and $P(K = e_3) = 0.4$.

Since $P(C = c \mid M = m)$ represents the probability that m gets enciphered as c, we can compute

$$P(C = c_1 \mid M = m_1) = P(K = e_1) \qquad P(C = c_1 \mid M = m_3) = P(K = e_3)$$
$$P(C = c_2 \mid M = m_1) = P(K = e_2) \qquad P(C = c_2 \mid M = m_3) = 0$$
$$P(C = c_3 \mid M = m_1) = P(K = e_3) \qquad P(C = c_3 \mid M = m_3) = P(K = e_1)$$
$$P(C = c_4 \mid M = m_1) = 0 \qquad P(C = c_4 \mid M = m_3) = P(K = e_2)$$

$$P(C = c_1 \mid M = m_2) = 0 \qquad P(C = c_1 \mid M = m_4) = P(K = e_2)$$
$$P(C = c_2 \mid M = m_2) = P(K = e_1) \qquad P(C = c_2 \mid M = m_4) = P(K = e_3)$$
$$P(C = c_3 \mid M = m_2) = P(K = e_2) \qquad P(C = c_3 \mid M = m_4) = 0$$
$$P(C = c_4 \mid M = m_2) = P(K = e_3) \qquad P(C = c_4 \mid M = m_4) = P(K = e_1)$$

For perfect secrecy, we require that $P(M \mid C) = P(M)$ or equivalently, $P(C \mid M) = P(C)$ for all $m \in M, c \in C$. However, since for example we have

$$P(C = c_1) = \sum_{m \in M} P(C = c_1 \mid M = m)P(M = m) = (0.3)(0.1) + (0.4)(0.3) + (0.3)(0.4) = 0.27$$

and $P(C = c_1 \mid M = m_1) = P(K = e_1) = 0.3$

It follows that we **cannot achieve** perfect secrecy in this case, since we do not have $P(C \mid M) = P(C)$ for all $m \in M, c \in C$ as is required in the definition of perfect secrecy.

Note that for this cryptosystem, we have $|M| = |C| > |K|$, so that by Theorem 7.4, we cannot have perfect secrecy regardless of the distributions of keys and messages.

Example 7.6 (The One-Time Pad): Consider the symmetric cryptosystem Γ with the message set $M = \{0,1\}^n$, cipher set $C = \{0,1\}^n$, and key set $K = \{e_1, e_2\}^n$ where the keys are defined as follows.

$$e_1: \quad 0 \to 1 \qquad e_2: \quad 0 \to 0$$
$$\qquad\quad 1 \to 0 \qquad\qquad\quad 1 \to 1$$

Note that the enciphering transformations in this case correspond to the Boolean XOR operation, i.e. $a \oplus a = 0, a \oplus b = 1$ for $a, b \in \{0,1\}$, where the key set acts on the message set one bit at a time.

Now suppose **A** wishes to send the binary message '00110100' to **B**, where **A** and **B** are assumed to share a private key. If **A** and **B** agree on using the sequence of keys '$e_2\ e_1\ e_2\ e_1\ e_2\ e_2\ e_2\ e_1$' as their private key, it is the same as adding the bits '01010001' via the Boolean XOR operation to the message sequence, resulting in the ciphertext '01100101' as indicated below.

$$\begin{array}{rl} 00110100 & (\textit{message}) \\ \oplus\ 01010001 & (\textit{key}) \\ \hline 01100101 & (\textit{cipher}) \end{array}$$

Then, since **B** also knows the private key, **B** decrypts by performing the Boolean XOR operation again as follows.

$$\begin{array}{rl} 01100101 & (\textit{cipher}) \\ \oplus\ 01010001 & (\textit{key}) \\ \hline 00110100 & (\textit{message}) \end{array}$$

It transpires that if, for each bit, the enciphering transformation is chosen at random (i.e. $P(K = e_i) = 1/2$ for all $e_i \in K$), then the one-time pad symmetric cryptosystem has perfect secrecy. In fact, we shall soon see that for a symmetric cryptosystem with $|M| = |C| = |K|$, then the cryptosystem has perfect secrecy **if and only if** the keys are equiprobable.

One significant drawback regarding the practical use of the one-time pad is the fact that the key bits are difficult to generate in a random fashion. Currently, the most widely used methods for generating the key bits 'pseudo-randomly' is by way of Linear Shift Feedback Registers (LFSRs) or by using linear congruences. Refer to [2] or [14] for more details.

It should be noted that in World War II, the one-time pad was used successfully, although once a given key was used, it had to be thrown out to ensure that if the cryptosystem was compromised, the message encrypted with that particular key would remain safe. This gave rise to the moniker, *one-time pad.*

7.2 Equiprobable Keys and Perfect Secrecy

We establish here a deeper connection between perfect secrecy and the general structure of the encryption keys. In addition to the fact that the number of keys must be greater than or equal to the number of messages as shown earlier, we present the following proposition which states that, as long as the set of messages, ciphers, and keys satisfy the equality $|M| = |C| = |K|$, then if the keys are distributed uniformly, perfect secrecy can be achieved, and vice versa.

Theorem 7.7: *For messages $m \in M$, ciphers $c \in C$, and keys $e_i \in K$ such that each $e_i \in K$ is an injective map from M to C, with $|M| = |C| = |K| = n$ where n is an integer, then $P(M \mid C) = P(M)$ if and only if each $e_i \in K$ is equiprobable.*

Proof: Assume that $|M| = |C| = |K| = n$. From our discussion above, we know that for each ordered pair of messages and ciphers (m, c), there exists a unique enciphering transformation such that $e_i(m_i) = c$ since $\{e_1(m_1), e_2(m_2), ..., e_n(m_n)\} = C$.

Therefore, we have that $P(C = c \mid M = m_i) = P(K = e_i)$. Now, since keys and messages are assumed to be independent of each other, then $P(M = m_i, k = e_j) = P(M = m_i)P(K = e_j)$.

From Bayes' Formula, we have

$$P(M = m_i \mid C = c) = \frac{P(M = m_i, C = c)}{P(C = c)} = \frac{P(M = m_i)P(K = e_i)}{\sum_{i=1}^{n} P(C = c \mid M = m_i)P(M = m_i)}$$

since the probability of having message m_i and cipher c is equivalent to using key e_i to encipher message m_i.

Upon applying the perfect secrecy condition and simplifying the right side further, we obtain

$$P(M = m_i) = \frac{P(M = m_i)P(K = e_i)}{P(M = m_1)P(K = e_1) + ... + P(M = m_n)P(K = e_n)} \qquad (7.6)$$

This gives $P(K = e_i) = P(M = m_1)P(K = e_1) + ... + P(M = m_n)P(K = e_n)$.

Furthermore, using the fact that $\sum_{i=1}^{n} P(M = m_i) = 1$, we have

$$P(K = e_i)[P(M = m_1) + ... + P(M = m_n)] = P(M = m_1)P(K = e_1) + ... + P(M = m_n)P(K = e_n)$$

which in turn implies

$$P(M = m_1)[P(K = e_i) - P(K = e_1)] + ... + P(M = m_n)[P(K = e_i) - P(K = e_n)] = 0$$

Now, choose the notation so that $P(K = e_i)$ is greater than or equal to $P(K = e_j)$ for any j, $1 \le j \le n$. Then, in the above, we have a sum of non-negative terms adding to zero. This can only happen if each term in the sum is zero.

Therefore, since $P(M = m_i) > 0$ for all $m_i \in M$, we have that $P(K = e_i) = \frac{1}{n}$ for all $e_i \in K$.

To show that equiprobable keys implies perfect secrecy, we can substitute $P(K = e_i) = \frac{1}{n}$ into (7.6) above to obtain the desired result immediately. Thus, we have that for a symmetric cryptosystem with $|M| = |C| = |K| = n$, where n is an integer, then $P(M \mid C) = P(M)$ if and only if each $e_i \in K$ is equiprobable.

□

7.3 Characterization of Perfect Secrecy

In the literature, it is frequently asserted that the only cryptosystem that exhibits perfect secrecy is the one-time pad. However, this assertion is false, and our efforts here will seek to dispel some of the apparent confusion by outlining a more complete characterization, particularly that perfect secrecy amounts to an n x n Latin square.

Suppose that $|M| = |C| = |K| = n$ where n is an integer. Also, let $M = \{m_1, m_2, ..., m_n\}$, $C = \{c_1, c_2, ..., c_n\}$, and $K = \{e_1, e_2, ..., e_n\}$. Since the keys are injective, we know that, for a fixed i, the set of encrypted messages $\{e_i(m_1), e_i(m_2), ..., e_i(m_n)\}$ is a set of distinct ciphers. If we consider our messages and ciphers to be the integers from 1 to n, i.e. $M = \{1,2,...,n\} = C$, then applying the i^{th} key to our message set produces a rearrangement of the set $\{1,2,...,n\}$. Finally, by varying i, we obtain all possible rearrangements of the messages. In matrix form, if each rearrangement represents a row, we can write this as

$$L = [e_{ij}] = \begin{bmatrix} e_1(1) & e_1(2) & \cdots & e_1(n) \\ e_2(1) & e_2(2) & \cdots & e_2(n) \\ \vdots & \vdots & \ddots & \vdots \\ e_n(1) & e_n(2) & \cdots & e_n(n) \end{bmatrix}.$$

where every row and column of L is a permutation of the set $\{1,2,...,n\}$ since the keys are injective and every message must be mapped to every cipher once and only once as a result of the condition that $|M| = |C| = |K|$. Such a construction is referred to as a *Latin square.* We now present a theorem from Bruen and Forcinito ([2]) that properly characterizes perfect secrecy when $|M| = |C| = |K|$.

Theorem 7.8 (Characterization of Perfect Secrecy): *Let* Γ *be a symmetric cryptosystem exhibiting perfect secrecy with* $|M| = |C| = |K| = n$*, so we may take M and C as the set of integers* $\{1,2,...,n\}$*. Each enciphering transformation* $e_i \in K$ *yields a unique row of an n x n Latin square L, i.e. a permutation of* $\{1,2,...,n\}$ *and the key is the index of that row. Each key is chosen with uniform probability. If the message is j and the enciphering key is* e_i*, we have* $e_i(j) = e_{ij}$*. Conversely, given any n x n Latin square L, we may construct a cryptosystem with perfect secrecy as above.*

Example 7.9: Suppose we have a symmetric cryptosystem Γ exhibiting perfect secrecy with $|M| = |C| = |K| = 4$, where $M = C = \{1,2,3,4\}$. Hence, we can represent Γ using a Latin square such as

$$L = \begin{bmatrix} 2 & 3 & 4 & 1 \\ 1 & 2 & 3 & 4 \\ 3 & 4 & 1 & 2 \\ 4 & 1 & 2 & 3 \end{bmatrix}$$

Now, suppose that **A** and **B** decide to use the fourth key, e_4. If **A** transmits the cipher '3' to **B**, then **B** searches the fourth row of L for '3'. Since '3' appears in the fourth column of the fourth row of L, **B** deduces that the original message was '4', i.e. '4' is the message m such that $e_4(m) = 3$.

As another example of the Latin square characterization of perfect secrecy, we can further support the craze of the popular Japanese game, Sudoku, which is based on filling in the entries of a 9 x 9 Latin square. However, it should be noted that for our purposes here, we are interested only in the larger square itself, and not the extra constraint that the nine disjoint sub-squares cannot have duplicate entries.

Example 7.10: Consider the following solved Sudoku puzzle, which is a 9 x 9 Latin square since the rows and columns are permutations of the set of integers $\{1,2,...,n\}$.

9	8	7	6	1	3	2	4	5
5	6	4	8	2	9	1	3	7
2	3	1	5	4	7	8	9	6
4	7	2	3	8	5	6	1	9
3	9	5	4	6	1	7	8	2
8	1	6	9	7	2	3	5	4
6	4	3	2	5	8	9	7	1
1	5	9	7	3	6	4	2	8
7	2	8	1	9	4	5	6	3

By Theorem 7.8, we can use the solved puzzle to construct a cryptosystem with perfect secrecy, where $M = \{1,..,9\} = C$, and the keys are the indexes of each row. Thus, we can see that the keys are defined as

e_1 :	e_2 :	...	e_9 :
1→ 9	1→ 5		1→ 7
2→ 8	2→ 6		2→ 2
3→ 7	3→ 4		3→ 8
4→ 6	4→ 8		4→ 1
5→ 1	5→ 2		5→ 9
6→ 3	6→ 9		6→ 4
7→ 2	7→ 1		7→ 5
8→ 4	8→ 3		8→ 6
9→ 5	9→ 7		9→ 3

Thus, if **A** and **B** agree to use key e_2, then the sequence '13579' would be encrypted as '54217', and since the keys are equiprobable, then **E**, who is eavesdropping on the transmission of the encrypted sequence, would be unable to determine which key was used to encipher the original message sequence. However, **B** would know to use the inverse of key e_2 to undo the encryption, or decrypt the sequence '54217' by determining the column in which each of the ciphers are found. Thus, since '5' is found in column 1 of the second row, then '5' decrypts to '1' and so on.

Therefore, we can conclude that the game of Sudoku amounts to determining how messages should be mapped to each cipher to ensure that perfect secrecy can be achieved!

As a final example, consider a cryptosystem with $M = \{0,1\} = C$, with the corresponding Latin square

$$L = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Since the keys are assumed to be uniformly distributed, we see that L actually corresponds to the channel transition matrix of a completely unreliable channel from M to C! This confirms our interpretation that perfect secrecy can be thought of as a discrete channel from M to C that has a capacity of zero, or equivalently, as a completely random channel. We also note

that the one-time pad discussed earlier gives rise to a Latin square of size 2^n x 2^n and so can be fitted into the characterization above.

7.4 Some New Examples of Perfect Secrecy

As we saw in the previous section, perfect secrecy can be attained by a cryptosystem as long as the messages, keys, and ciphers satisfy the inequality $|M| \le |C| \le |K|$. In the literature, there are examples of cryptosystems exhibiting perfect secrecy for the case where $|M| = |C| = |K|$ such as the one-time pad, but few others are discussed. To address these shortcomings, we have already introduced some new examples using Latin squares, and we now present some new examples of cryptosystems achieving perfect secrecy with $|M| < |C| < |K|$.

***Example 7.11*:** Consider the discrete cryptosystem with $M = \{0,1\}$, $C = \{0,1,2\}$, and $K = \{e_1, e_2, e_3, e_4\}$, where $P(M = 0) = p, P(M = 1) = 1 - p$. In order to satisfy the perfect secrecy condition, we must have $P(C = c \mid M = m) = P(C = c)$ for all $c \in C, m \in M$.
By the Law of Total Probability, we have $P(C = c) = \sum_{m \in M} P(C = c \mid M = m) P(M = m)$,

thus, in order to satisfy $P(C = c \mid M = m) = P(C = c)$, we need for each $c \in C$,

i. $P(C = c) = P(C = c \mid M = 0)$ and
ii. $P(C = c) = P(C = c \mid M = 1)$

For $C = 0$, we obtain $P(C = 0) = p \cdot P(C = 0 \mid M = 0) + (1 - p) \cdot P(C = 0 \mid M = 1)$ using the Law of Total Probability. Similarly, we compute

$$P(C = 1) = p \cdot P(C = 1 \mid M = 0) + (1 - p) \cdot P(C = 1 \mid M = 1)$$
$$P(C = 2) = p \cdot P(C = 2 \mid M = 0) + (1 - p) \cdot P(C = 2 \mid M = 1)$$

So, for (i), we need to choose our keys to simultaneously satisfy the following requirements:

$$p \cdot P(C=0 \mid M=0) + (1-p) \cdot P(C=0 \mid M=1) = P(C=0 \mid M=0)$$
$$p \cdot P(C=1 \mid M=0) + (1-p) \cdot P(C=1 \mid M=1) = P(C=1 \mid M=0)$$
$$p \cdot P(C=2 \mid M=0) + (1-p) \cdot P(C=2 \mid M=1) = P(C=2 \mid M=0)$$

Equivalently, we need to choose our set of keys K such that

$$\begin{aligned} P(C=0 \mid M=0) &= P(C=0 \mid M=1) \\ P(C=1 \mid M=0) &= P(C=1 \mid M=1) \\ P(C=2 \mid M=0) &= P(C=2 \mid M=1) \end{aligned} \tag{7.7}$$

Then, as long as the keys are constructed in such a way that each message is mapped to a given cipher the same number of times, we can achieve perfect secrecy. Now, consider the set of keys defined as follows.

e_1:	$0 \to 1$	e_2:	$0 \to 2$	e_3:	$0 \to 0$	e_4:	$0 \to 1$
	$1 \to 2$		$1 \to 0$		$1 \to 1$		$1 \to 0$

where $P(K = e_i) \geq 0$ for all $e_i \in K$. Then using these keys to encrypt our message set M, we obtain the following conditional probabilities in terms of key probabilities.

$$P(C=0 \mid M=0) = P(K=e_3) \qquad P(C=0 \mid M=1) = P(K=e_2) + P(K=e_4)$$
$$P(C=1 \mid M=0) = P(K=e_1) + P(K=e_4) \qquad P(C=1 \mid M=1) = P(K=e_3)$$
$$P(C=2 \mid M=0) = P(K=e_2) \qquad P(C=2 \mid M=1) = P(K=e_1)$$

Hence, to satisfy the system of equations from above, we require that

i. $P(K=e_3)=P(K=e_2)+P(K=e_4)$

ii. $P(K=e_1)=P(K=e_1)$

iii. $P(K=e_1)+P(K=e_2)+P(K=e_3)+P(K=e_4)=1$

where (iii) comes from the fact that the key probabilities belong to a probability distribution.

In matrix form, the above system can be represented as the row reduced matrix

$$\begin{bmatrix} 0 & 1 & -1 & 1 & 0 \\ 1 & 0 & -1 & 1 & 0 \\ 1 & -1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 0 & 0 & \frac{2}{3} & \frac{1}{3} \\ 0 & 1 & 0 & \frac{2}{3} & \frac{1}{3} \\ 0 & 0 & 1 & \frac{-1}{3} & \frac{1}{3} \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

where '$\Rightarrow$' denotes row-reduction.

This results in the set of solutions given by

$$X = \begin{bmatrix} P(K=e_1) \\ P(K=e_2) \\ P(K=e_3) \\ P(K=e_4) \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \\ 0 \end{bmatrix} + t \begin{bmatrix} -2/3 \\ -2/3 \\ 1/3 \\ 1 \end{bmatrix}$$

with t such that $P(K=e_i) \geq 0$ for all $e_i \in K$.

Hence, the cryptosystem with $M=\{0,1\}, C=\{0,1,2\}, K=\{e_1,e_2,e_3,e_4\}$ where the keys are defined as above, exhibits perfect secrecy as long as the keys have a probability distribution given by

$$\{P(K=e_i)\}=\left\{\frac{1-2t}{3},\frac{1-2t}{3},\frac{1+t}{3},t\right\}$$

which provides us with an infinite number of possibilities for key distributions!

For another example, we can make use of the same sets of messages and ciphers, while increasing the number of possible keys.

Example 7.12: Consider the discrete cryptosystem with $M=\{0,1\}, C=\{0,1,2\}$, and $K=\{e_1,e_2,e_3,e_4,e_5,e_6\}$, where $P(M=0)=p, P(M=1)=1-p$. Again, in order to satisfy the perfect secrecy condition, we must have $P(C=c\,|\,M=m)=P(C=c)$ for all $c\in C, m\in M$.

As was the case in Example 7.4, for perfect secrecy we must choose our keys to satisfy

$$\begin{aligned} P(C=0\,|\,M=0)&=P(C=0\,|\,M=1)\\ P(C=1\,|\,M=0)&=P(C=1\,|\,M=1)\\ P(C=2\,|\,M=0)&=P(C=2\,|\,M=1) \end{aligned} \qquad (7.8)$$

Let our keys be defined in the following way, again with $P(K=e_i)\geq 0$ for all $e_i\in K$.

e_1:	$0\to 0$	e_2:	$0\to 0$	e_3:	$0\to 1$	e_4:	$0\to 1$	e_5:	$0\to 2$	e_6:	$0\to 2$
	$1\to 1$		$1\to 2$		$1\to 0$		$1\to 2$		$1\to 0$		$1\to 1$

Thus we can determine that

$$\begin{aligned} P(C=0\,|\,M=0)&=P(K=e_1)+P(K=e_2) & P(C=0\,|\,M=1)&=P(K=e_3)+P(K=e_5)\\ P(C=1\,|\,M=0)&=P(K=e_3)+P(K=e_4) & P(C=1\,|\,M=1)&=P(K=e_1)+P(K=e_6)\\ P(C=2\,|\,M=0)&=P(K=e_5)+P(K=e_6) & P(C=2\,|\,M=1)&=P(K=e_2)+P(K=e_4) \end{aligned}$$

and, along with the fact that $P(K=e_1)+P(K=e_2)+P(K=e_3)+P(K=e_4)=1$, we can solve the system of equations for key probabilities to obtain the row reduced matrix

$$\begin{bmatrix} 1 & 1 & -1 & 0 & -1 & 0 & 0 \\ 1 & 0 & -1 & -1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & -1 & -1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 0 & 0 & \frac{-1}{2} & 1 & \frac{3}{2} & \frac{1}{2} \\ 0 & 1 & 0 & 1 & -1 & -1 & 0 \\ 0 & 0 & 1 & \frac{1}{2} & 1 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

where '$\Rightarrow$' denotes row-reduction.

Based on the row-reduced matrix, we arrive at the of set of solutions

$$X = \begin{bmatrix} P(K=e_1) \\ P(K=e_2) \\ P(K=e_3) \\ P(K=e_4) \\ P(K=e_5) \\ P(K=e_6) \end{bmatrix} = \begin{bmatrix} 1/2 \\ 0 \\ 1/2 \\ 0 \\ 0 \\ 0 \end{bmatrix} + s\begin{bmatrix} 1/2 \\ -1 \\ -1/2 \\ 1 \\ 0 \\ 0 \end{bmatrix} + t\begin{bmatrix} -1 \\ 1 \\ -1 \\ 0 \\ 1 \\ 0 \end{bmatrix} + u\begin{bmatrix} -3/2 \\ 1 \\ -1/2 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

with s, t, and u such that $P(K=e_i)\geq 0$ for all $e_i \in K$.

To see that there is at least one valid solution, with $s=\frac{1}{6}$, $t=\frac{1}{6}$, and $u=\frac{1}{18}$, we obtain

$$X = \begin{bmatrix} P(K=e_1) \\ P(K=e_2) \\ P(K=e_3) \\ P(K=e_4) \\ P(K=e_5) \\ P(K=e_6) \end{bmatrix} = \begin{bmatrix} 6/18 \\ 1/18 \\ 4/18 \\ 3/18 \\ 3/18 \\ 1/18 \end{bmatrix}$$

Contrast both examples with the case when $|M| = |C| = |K|$ where the keys must be uniformly distributed in order to achieve perfect secrecy. By increasing the number of keys and ciphers, we have gained some flexibility in how often each key must be used without sacrificing perfect secrecy.

7.5 Perfect Secrecy for Analog Communication Systems

As we saw in Chapter 6, Shannon's Sampling Theorem provides the framework for which efficient analog communications is made possible. What isn't widely known is the fact that the same theorem can be used as the basis for an analog, symmetric cryptosystem which offers unconditional security much like the one-time pad. The idea was first proposed by Alan Turing during the later stages of World War II, and combines the concept of the one-time pad with sampled analog signals. Turing's original objective was to produce a secure speech encoder for the British, although his cryptosystem is well-suited for any short range analog applications.

Turing's cryptosystem, referred to as "Delilah" (or 'deceiver of men' in Biblical times), acts on a time-varying continuous signal, for which samples are taken according to the Nyquist rate. Without loss of generality, we will assume here that the signal $x(t)$ is positive for all values of t.

Hence, upon sampling at a rate of $T_s = 1/2W$ where W is the bandwidth of the original signal, a set of samples $\{x(nT_s)\}_{n=0}^{N}$ is produced, where N is the number of samples, and n is an integer, as shown below in Figure 7-1.

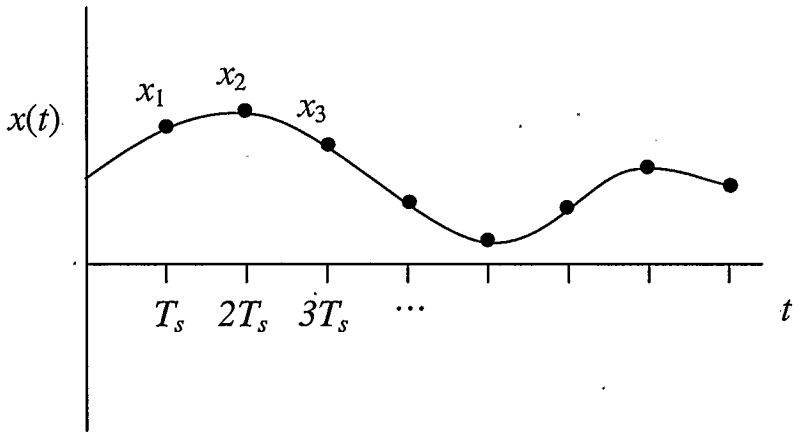

Figure 7-1 A signal and its samples taken at the Nyquist rate

Then, using a pseudo-random number generator such as a Linear Feedback Shift Register (LFSRs, as discussed in [2]), N key values can be generated and added to the sampled values modulo A, where A is the amplitude of the original signal as shown in Figure 7-2.

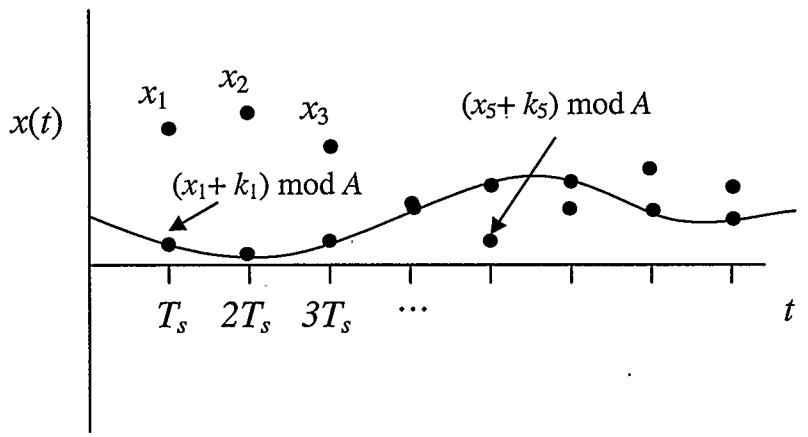

Figure 7-2 An encrypted signal and its samples

The resulting set of sample values is then transmitted using standard analog communication techniques (i.e. AM or FM modulation), and upon reception, the receiver then subtracts each key value from the corresponding samples, and reconstructs the original signal using the reconstruction formula

$$x(t) = \sum_{n=-\infty}^{\infty} 2W_1 T_s x(nT_s)\text{sinc}(2W_1(t - nT_s)) \qquad (7.9)$$

where W_1 is any number such that $W \leq W_1 < f_s - W$ and $f_s \leq 2W$ from Chapter 6.

Since the keys are ideally generated in a random fashion, each key is equally likely and hence, each message is equally likely. Moreover, since each key value is a real number, the number of possible keys and in turn, the number of possible messages is uncountable. Therefore, Delilah indeed provides perfect secrecy under ideal conditions for the same reason the one-time pad does. The only foreseeable drawback with Delilah, in addition to the drawbacks discussed with regards to the one-time pad, is the fact that as with analog communication in general, a small amount of signal perturbation or time delay significantly impacts the fidelity of the transmitted signal. Hence, Delilah's effective range is quite short relative to other, less secure analog encryption schemes such as the X–System[1]. However, depending on the particular application, the theoretically unconditional security offered by Delilah may far outweigh the advantage of extended useable range offered by other analog encryption systems.

[1] American speech encryption scheme (refer to [10]).

8 Discussion & Conclusions

We have provided here a wide-ranging yet focused presentation regarding the applications of Information Theory in the fields of communication channels, signal processing, and cryptography. We have presented several new results pertaining to discrete channels, including a generalization of Shannon's theorem for regular channels, which allows for a significant reduction in the calculation of channel capacity for channels that are near-regular. We have also cleared up some discrepancies and omissions in the literature regarding the computation of channel capacity, outlining several new techniques that can be used to solve such problems.

We also presented here an axiomatization of certain reciprocal channels, along with some new developments, including the computation of the capacity of the reciprocal binary symmetric channel using the symmetry of mutual information. Reciprocal channels have proven to be very useful in practice, and our efforts here may well help simplify the analysis and characterization of channels in the future.

In the realm of cryptography, the proper characterization of perfect secrecy was outlined in the form of Latin squares, contradicting the impression or assertion in the standard literature that perfect secrecy must amount to the one-time pad when the number of messages, ciphers, and keys are the same. We also provided some new examples of perfect secrecy when the number of messages, ciphers, and keys are not equal. Finally, Turing's proposed cryptosystem for analog signals based on the sampling theorem was discussed, showing that analog signals should not be ignored when it comes to cryptographic applications, as is currently the case for the most part.

With regards to future work in the area of Information Theory, we have only scratched the surface in terms of the applicability of Shannon's ideas. In the realm of communication theory, much work has been done over the past several years extending Information Theory to multiple input, multiple output (MIMO) channels and continuous channels subject to other types of distortion such as Rayleigh fading. Moreover, Information Theory has proven to be

fundamental in the analysis of more leisurely pursuits such as poker and horse-racing, via the Kelly formula ([16]), as well as stock-market analysis, molecular biology ([2]), and jurisprudence ([11]). In short, the applications of Information Theory extend far beyond the reaches of communication theory, giving credence to the argument, in reference to Shannon's revolutionary paper, *A Mathematical Theory of Communication* ([19]), that "no other works of the twentieth century have had greater impacts on science and engineering."([2])

Bibliography

[1] Ash, R.B. *Information Theory.* 1st Edition, Dover Books, 1990.

[2] Bruen, A.A. and M. Forcinito. *Cryptography, Information Theory and Error Correction: A Handbook for the 21st Century.* 1st Edition, John Wiley & Sons, New York, 2005.

[3] Cohen, L. *Time-Frequency Analysis.* 1st Edition, Prentice Hall, New Jersey, 1995.

[4] Cover, T. and J. Thomas. *Elements of Information Theory.* 1st Edition, John Wiley & Sons, New York, 1991.

[5] Davis, J. *Methods of Applied Mathematics.* 1st Edition, Queen's University, Kingston, 2000.

[6] Gersho, A. and R.M. Gray. *Vector Quantization and Signal Compression.* 9th Edition, Kluwer Academic Publishers, Boston, 2003.

[7] Ghahramani, S. *Fundamentals of Probability.* 2nd Edition, Prentice Hall, New Jersey, 2000.

[8] Goldie, M.C. & R.G.E. Pinch. *Communication Theory.* 1st Edition, Cambridge University Press, 1991.

[9] Hamming, R.W. *Digital Filters.* 3rd Edition, Dover Publications Inc., New York, 1989.

[10] Hodges, A. *Alan Turing: The Enigma.* 5th Edition, Vintage, London, 1992.

[11] Klarreich, E. *Ideal Justice*. Science News, Volume 163, pg 405, 2003.

[12] Larson, R., Hostetler, R.P. & B.H. Edwards. *Multivariable Calculus*. 8th Edition, Houghton Mifflin, New York, 2006.

[13] R.P. Lathi. *Modern Digital and Analog Communication Systems*. 3rd Edition, Oxford University Press, London, 1998.

[14] Luenberger, D.G. *Information Science*. 1st Edition, Princeton University Press, New Jersey, 2006.

[15] Nielsen, M.A. and I.L. Chuang. *Quantum Computation & Quantum Information*. 1st Edition, Cambridge University Press, 2000.

[16] Poundstone, W. *Fortune's Formula: the untold story of the scientific betting system that beat the casinos and Wall Street*. 1st Edition, Hill and Wang, New York, 2005.

[17] Proakis, J.G., and M. Salehi. *Communication Systems Engineering*. 2nd Edition, Prentice Hall, New Jersey, 2002.

[18] Ross, S.M. *Introduction to Probability Models*. 8th Edition, Academic Press, New York, 2003.

[19] Shannon, C.E. *A Mathematical Theory of Communication*. Bell Systems Technical Journal, Volume 27, pp: 379-423, 623-656, 1948.

[20] Shannon, C.E. *Communication Theory of Secrecy Systems*. Bell Systems Technical Journal, Volume 28, pp: 656-715, 1949.

[21] Welsh, D. *Codes and Cryptography*. 1st Edition, Oxford University Press, 1988.

Appendix A

The Maple calculations pertaining to reciprocal channels in Chapter 5 are provided here.

Ash's Formula:

```
> u:=(q2*(-q1*log[2](q1)-p1*log[2](p1))-p1*(-p2*log[2](p2)-
q2*log[2](q2)))/(p2-q1);
```

$$u := \frac{q2\left(-\frac{q1\ln(q1)}{\ln(2)}-\frac{p1\ln(p1)}{\ln(2)}\right)-p1\left(-\frac{p2\ln(p2)}{\ln(2)}-\frac{q2\ln(q2)}{\ln(2)}\right)}{p2-q1}$$

```
> v:=(-p2*(-q1*log[2](q1)-p1*log[2](p1))+q1*(-p2*log[2](p2)-
q2*log[2](q2)))/(p2-q1);
```

$$v := \frac{-p2\left(-\frac{q1\ln(q1)}{\ln(2)}-\frac{p1\ln(p1)}{\ln(2)}\right)+q1\left(-\frac{p2\ln(p2)}{\ln(2)}-\frac{q2\ln(q2)}{\ln(2)}\right)}{p2-q1}$$

```
> Cap:=log[2](2^u+2^v);
```

$$Cap := \ln\left(2^{\left(\frac{q2\left(-\frac{q1\ln(q1)}{\ln(2)}-\frac{p1\ln(p1)}{\ln(2)}\right)-p1\left(-\frac{p2\ln(p2)}{\ln(2)}-\frac{q2\ln(q2)}{\ln(2)}\right)}{p2-q1}\right)}+2^{\left(\frac{-p2\left(-\frac{q1\ln(q1)}{\ln(2)}-\frac{p1\ln(p1)}{\ln(2)}\right)+q1\left(-\frac{p2\ln(p2)}{\ln(2)}-\frac{q2\ln(q2)}{\ln(2)}\right)}{p2-q1}\right)}\right)/\ln(2)$$

```
>
Cap1:=subs([q1=(a*q)/(a*q+b*p),p1=(a*p)/(a*p+b*q),p2=(b*p)/(a*
q+b*p),q2=(b*q)/(a*p+b*q)],Cap);
```

$$Cap1 := \ln\left(2^{\left(\frac{bq\left(-\frac{aq\ln\left(\frac{aq}{aq+bp}\right)}{(aq+bp)\ln(2)}-\frac{ap\ln\left(\frac{ap}{ap+bq}\right)}{(ap+bq)\ln(2)}\right)}{ap+bq}-\frac{ap\left(-\frac{bp\ln\left(\frac{bp}{aq+bp}\right)}{(aq+bp)\ln(2)}-\frac{bq\ln\left(\frac{bq}{ap+bq}\right)}{(ap+bq)\ln(2)}\right)}{ap+bq}}{\frac{bp}{aq+bp}-\frac{aq}{aq+bp}}\right)}+2^{\left(\frac{-\frac{bp\left(-\frac{aq\ln\left(\frac{aq}{aq+bp}\right)}{(aq+bp)\ln(2)}-\frac{ap\ln\left(\frac{ap}{ap+bq}\right)}{(ap+bq)\ln(2)}\right)}{aq+bp}+\frac{aq\left(-\frac{bp\ln\left(\frac{bp}{aq+bp}\right)}{(aq+bp)\ln(2)}-\frac{bq\ln\left(\frac{bq}{ap+bq}\right)}{(ap+bq)\ln(2)}\right)}{aq+bp}}{\frac{bp}{aq+bp}-\frac{aq}{aq+bp}}\right)}\right)/\ln(2)$$

```
> Cap2:=subs(q=1-p,Cap1);
```

$$Cap2 := \ln\left(2^{\left(\left(\frac{b(1-p)\left(-\frac{a(1-p)\ln\left(\frac{a(1-p)}{a(1-p)+bp}\right)}{(a(1-p)+bp)\ln(2)}-\frac{ap\ln\left(\frac{ap}{ap+b(1-p)}\right)}{(ap+b(1-p))\ln(2)}\right)}{ap+b(1-p)}\right.\right.}\right.$$

$$\left.\left.-\frac{ap\left(-\frac{bp\ln\left(\frac{bp}{a(1-p)+bp}\right)}{(a(1-p)+bp)\ln(2)}-\frac{b(1-p)\ln\left(\frac{b(1-p)}{ap+b(1-p)}\right)}{(ap+b(1-p))\ln(2)}\right)}{ap+b(1-p)}\right)\Big/\left(\frac{bp}{a(1-p)+bp}-\frac{a(1-p)}{a(1-p)+bp}\right)\right)+2^{\left(\left(\right.\right.}$$

$$-\frac{bp\left(-\frac{a(1-p)\ln\left(\frac{a(1-p)}{a(1-p)+bp}\right)}{(a(1-p)+bp)\ln(2)}-\frac{ap\ln\left(\frac{ap}{ap+b(1-p)}\right)}{(ap+b(1-p))\ln(2)}\right)}{a(1-p)+bp}$$

$$\left.\left.+\frac{a(1-p)\left(-\frac{bp\ln\left(\frac{bp}{a(1-p)+bp}\right)}{(a(1-p)+bp)\ln(2)}-\frac{b(1-p)\ln\left(\frac{b(1-p)}{ap+b(1-p)}\right)}{(ap+b(1-p))\ln(2)}\right)}{a(1-p)+bp}\right)\Big/\left(\frac{bp}{a(1-p)+bp}-\frac{a(1-p)}{a(1-p)+bp}\right)\right)\Big)/$$

$$\ln(2)$$

```
> Capacity:=simplify(Cap2);
```

$$Capacity := \ln\left(e^{\left(-ba\left(\ln\left(\frac{a(-1+p)}{-a+ap-bp}\right)ap+\ln\left(\frac{a(-1+p)}{-a+ap-bp}\right)b-3\ln\left(\frac{a(-1+p)}{-a+ap-bp}\right)bp\right.\right.}\right.$$

$$-2\ln\left(\frac{a(-1+p)}{-a+ap-bp}\right)ap^2+3\ln\left(\frac{a(-1+p)}{-a+ap-bp}\right)bp^2+ap\ln\left(\frac{pa}{ap+b-bp}\right)-2p^2\ln\left(\frac{pa}{ap+b-bp}\right)a$$

$$+p^2\ln\left(\frac{pa}{ap+b-bp}\right)b+\ln\left(\frac{a(-1+p)}{-a+ap-bp}\right)ap^3-\ln\left(\frac{a(-1+p)}{-a+ap-bp}\right)bp^3+p^3\ln\left(\frac{pa}{ap+b-bp}\right)a$$

$$-p^3\ln\left(\frac{pa}{ap+b-bp}\right)b-p^3\ln\left(-\frac{bp}{-a+ap-bp}\right)a-p^2\ln\left(-\frac{bp}{-a+ap-bp}\right)b+p^3\ln\left(-\frac{bp}{-a+ap-bp}\right)b$$

$$-\ln\left(-\frac{b(-1+p)}{ap+b-bp}\right)ap+2\ln\left(-\frac{b(-1+p)}{ap+b-bp}\right)ap^2-\ln\left(-\frac{b(-1+p)}{ap+b-bp}\right)bp^2-\ln\left(-\frac{b(-1+p)}{ap+b-bp}\right)ap^3$$

$$\left.+\ln\left(-\frac{b(-1+p)}{ap+b-bp}\right)bp^3\right)\Big/((ap+b-bp)^2(bp-a+ap))\Big)+e^{\left(ba\left(-\ln\left(\frac{a(-1+p)}{-a+ap-bp}\right)ap^2\right.\right.}$$

$$-\ln\left(\frac{a(-1+p)}{-a+ap-bp}\right)bp+2\ln\left(\frac{a(-1+p)}{-a+ap-bp}\right)bp^2+\ln\left(\frac{a(-1+p)}{-a+ap-bp}\right)ap^3-\ln\left(\frac{a(-1+p)}{-a+ap-bp}\right)bp^3$$

$$-p^2\ln\left(\frac{pa}{ap+b-bp}\right)a+p^3\ln\left(\frac{pa}{ap+b-bp}\right)a-p^3\ln\left(\frac{pa}{ap+b-bp}\right)b+p^2\ln\left(-\frac{bp}{-a+ap-bp}\right)a$$

$$+bp\ln\left(-\frac{bp}{-a+ap-bp}\right)-2p^2\ln\left(-\frac{bp}{-a+ap-bp}\right)b+\ln\left(-\frac{b(-1+p)}{ap+b-bp}\right)a-3\ln\left(-\frac{b(-1+p)}{ap+b-bp}\right)ap$$

$$+\ln\left(-\frac{b(-1+p)}{ap+b-bp}\right)bp+3\ln\left(-\frac{b(-1+p)}{ap+b-bp}\right)ap^2-2\ln\left(-\frac{b(-1+p)}{ap+b-bp}\right)bp^2-p^3\ln\left(-\frac{bp}{-a+ap-bp}\right)a$$

$$+p^3\ln\left(-\frac{bp}{-a+ap-bp}\right)b-\ln\left(-\frac{b(-1+p)}{ap+b-bp}\right)ap^3+\ln\left(-\frac{b(-1+p)}{ap+b-bp}\right)bp^3\Big)\Big/((-a+ap-bp)(ap+b-bp)$$

$$\left.(bp-a+ap))\right)\Big)/\ln(2)$$

Note that the above formula involves both input probabilities for the channel in addition to the parameter p.

Lagrange Multipliers:

```
> HY=-a*log[2](a)-b*log[2](b);
```

$$HY = -\frac{a\ln(a)}{\ln(2)} - \frac{b\ln(b)}{\ln(2)}$$

```
> HYX:=-(a-a*p)*log[2]((a-a*p)/(a-a*p+b*p))-
(p*b)*log[2]((p*b)/(a-a*p+b*p))-(p*a)*log[2]((p*a)/(p*a+b-
b*p))-(b-b*p)*log[2]((b-b*p)/(a*p+b-b*p));
```

$$HYX := -\frac{(a-ap)\ln\left(\frac{a-ap}{a-ap+bp}\right)}{\ln(2)} - \frac{pb\ln\left(\frac{pb}{a-ap+bp}\right)}{\ln(2)} - \frac{pa\ln\left(\frac{pa}{ap+b-bp}\right)}{\ln(2)} - \frac{(b-bp)\ln\left(\frac{b-bp}{ap+b-bp}\right)}{\ln(2)}$$

```
> with(Student[MultivariateCalculus]):
> LagrangeMultipliers(HY-HYX,[a+b-1],[a,b]);
```

$$\left[\begin{array}{l}
\mathrm{RootOf}\left(\mathrm{e}\right. \\
\quad -p\ln\left(-\frac{p^2(\mathrm{e}^{-Z}-1)}{-2\,\mathrm{e}^{-Z}p+\mathrm{e}^{-Z}+p^2}\right)+\ln\left(-\frac{-2p+1+p^2}{(-2\,\mathrm{e}^{-Z}p+\mathrm{e}^{-Z}+p^2)(\mathrm{e}^{-Z}-1)}\right)+_Z-p\ln\left(-\frac{-2p+1+p^2}{(-2\,\mathrm{e}^{-Z}p+\mathrm{e}^{-Z}+p^2)(\mathrm{e}^{-Z}-1)}\right)\Big) \\
(-1+p)\Big/\left(2\,\mathrm{e}^{\mathrm{RootOf}\left(\right.}\right. \\
\quad -p\ln\left(-\frac{p^2(\mathrm{e}^{-Z}-1)}{-2\,\mathrm{e}^{-Z}p+\mathrm{e}^{-Z}+p^2}\right)+\ln\left(-\frac{-2p+1+p^2}{(-2\,\mathrm{e}^{-Z}p+\mathrm{e}^{-Z}+p^2)(\mathrm{e}^{-Z}-1)}\right)+_Z-p\ln\left(-\frac{-2p+1+p^2}{(-2\,\mathrm{e}^{-Z}p+\mathrm{e}^{-Z}+p^2)(\mathrm{e}^{-Z}-1)}\right)\Big) \\
p-\mathrm{e}^{\mathrm{RootOf}\left(\right.} \\
\quad -p\ln\left(-\frac{p^2(\mathrm{e}^{-Z}-1)}{-2\,\mathrm{e}^{-Z}p+\mathrm{e}^{-Z}+p^2}\right)+\ln\left(-\frac{-2p+1+p^2}{(-2\,\mathrm{e}^{-Z}p+\mathrm{e}^{-Z}+p^2)(\mathrm{e}^{-Z}-1)}\right)+_Z-p\ln\left(-\frac{-2p+1+p^2}{(-2\,\mathrm{e}^{-Z}p+\mathrm{e}^{-Z}+p^2)(\mathrm{e}^{-Z}-1)}\right)\Big) \\
-p\Big), p\left(\mathrm{e}^{\mathrm{RootOf}\left(\right.}\right. \\
\quad -p\ln\left(-\frac{p^2(\mathrm{e}^{-Z}-1)}{-2\,\mathrm{e}^{-Z}p+\mathrm{e}^{-Z}+p^2}\right)+\ln\left(-\frac{-2p+1+p^2}{(-2\,\mathrm{e}^{-Z}p+\mathrm{e}^{-Z}+p^2)(\mathrm{e}^{-Z}-1)}\right)+_Z-p\ln\left(-\frac{-2p+1+p^2}{(-2\,\mathrm{e}^{-Z}p+\mathrm{e}^{-Z}+p^2)(\mathrm{e}^{-Z}-1)}\right)\Big) \\
-1\Big)\Big/\left(2\,\mathrm{e}^{\mathrm{RootOf}\left(\right.}\right. \\
\quad -p\ln\left(-\frac{p^2(\mathrm{e}^{-Z}-1)}{-2\,\mathrm{e}^{-Z}p+\mathrm{e}^{-Z}+p^2}\right)+\ln\left(-\frac{-2p+1+p^2}{(-2\,\mathrm{e}^{-Z}p+\mathrm{e}^{-Z}+p^2)(\mathrm{e}^{-Z}-1)}\right)+_Z-p\ln\left(-\frac{-2p+1+p^2}{(-2\,\mathrm{e}^{-Z}p+\mathrm{e}^{-Z}+p^2)(\mathrm{e}^{-Z}-1)}\right)\Big)
\end{array}\right]$$

$$\text{RootOf}\left(p - e^{-p\ln\left(-\frac{p^2(e^{-Z}-1)}{-2e^{-Z}p+e^{-Z}+p^2}\right)+\ln\left(-\frac{-2p+1+p^2}{(-2e^{-Z}p+e^{-Z}+p^2)(e^{-Z}-1)}\right)+\_Z-p\ln\left(-\frac{-2p+1+p^2}{(-2e^{-Z}p+e^{-Z}+p^2)(e^{-Z}-1)}\right)}-p\right)\right]$$

The two expressions above pertain to the maximizing values of $P(Y=0)$ and $P(Y=1)$. Notice that they both depend on the parameter p, which prevents us from obtaining a closed form for capacity.